Basics of Bayesian learning – Basically Bayes

Jan Larsen

Intelligent Signal Processing Group
Informatics and Mathematical Modelling
Technical University of Denmark

www.imm.dtu.dk
jl@imm.dtu.dk

DTU

Bayesian learning

T. Bayes "An Essay Towards Solving a Problem in the Doctrine of Chances, *Phil. Trans. Roy. Soc.*, **53**, 370–418, 1783

- The world is uncertain.....

  inference: assign probabilities to hypothesis from specific data set

  decision theory: choose between actions to minimize loss/risk

- Basic axiom systems for decision theory and inference leads to that rational analysis must corresponds to a Bayesian paradigm [Berger]
- You are probably already doing Bayes – even if you don't know it

Intention

- Crash course in Bayesian learning for those unfamiliar with this paradigm
- Experienced people hopefully gets new inspiration

DTU

Outline

- Why Bayesian learning?
- Basic ingredients
- Bayes estimators
- More on selection of priors
- Generalization and bias/variance
- Generalization estimation
- Bayesian model selection
- Discussion of Bayesian framework
- Example of Bayesian learning: RVM
- Bayesian signal detection

General Resources

- D. MacKay: *Information Theory, Inference and Learning Algorithms*, http://www.inference.phy.cam.ac.uk/mackay/itprnn
- J.O. Berger: *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, 2nd edition, 1985.
- C.P. Robert: *The Bayesian Choice: A Decision-Theoretic Motivation*, Springer-Verlag, 1994.
- J.J.K. Ó Ruanaidh and W.J. Fitzgerald: *Numerical Bayesian Methods Applied to Signal Processing*, Spinger-Verlag, 1996.

Outline

- Why Bayesian learning?
- Basic ingredients
- Bayes estimators
- More on selection of priors
- Generalization and bias/variance
- Generalization estimation
- Bayesian model selection
- Discussion of Bayesian framework
- Example of Bayesian learning: RVM
- Bayesian signal detection

Why Bayesian learning?

- principal framework which combines available uncertain knowledge - data, prior etc.
- Bayesian learning is optimal - if you are
- Bayesian learning is typically more robust to mis-specifications and small data sets
- classical learning schemes are special cases
- known to give better performance for most models
- offers model selection as an integrated part

Why Bayesian learning?

- predictions/forecast comes with errorbars (credible sets, highest posterior density credibility set)
- new approaches such as Variational Bayes, Expectation Propagation makes the Bayesian learning computational attractive

Applications of Bayesian learning

- clustering using mixture of Gaussians (Attias, Rasmussen)
- mixture of factor analyzers clustering and dimensionality reduction (Ghahramani+Beal)
- principal components analysis (Bishop)
- independent component analysis (Højen-Sørensen+Whinter+Hansen, Lee, Attais, Valpola, Miskin+MacKay)
- state-space models, e.g., extended Kalman filters (Ghahramani+Beal, de Freitas, Niranjan, Wan, Doucet, Gordon)
- time series modeling (Roberts+Penny, Quiñonero+Girard+Larsen+Rasmussen)

Applications of Bayesian learning (cont.)

- mixture of experts (Ueda)
- hidden Markov models (MacKay)
- Bayesian networks, graphical models (Heckerman, Jordan, Ghahramani, Bishop, Spiegelhalter)

Software Tools

- VIBES (Bishop, Spiegelhalter, Winn) http://vibes.sourceforge.net/
- Matthew Beal http://www.cse.buffalo.edu/faculty/mbeal/software.html
- ICA toolbox (DTU) http://isp.imm.dtu.dk/toolbox
- Bayes Blocks (HUT) http://www.cis.hut.fi/projects/bayes/software
- Bayes Net toolbox (Kevin Murphy) http://bnt.sourceforge.net
- ReBEL : Recursive Bayesian Estimation Library (E. Wan) http://choosh.ece.ogi.edu/rebel
- NetLab (Bishop) http://www.ncrg.aston.ac.uk/netlab/index.php

Basic ideas of Bayesian framework

- all variables have associated probability densities
- variables not required in the final estimate are integrated out

Outline

- Why Bayesian learning?
- Basic ingredients
- Bayes estimators
- More on selection of priors
- Generalization and bias/variance
- Generalization estimation
- Bayesian model selection
- Discussion of Bayesian framework
- Example of Bayesian learning: RVM
- Bayesian signal detection

The basic ingredients

- variables which we want to infer
- problems (unsupervised/supervised)
- data
- model
- prior
- predictive distribution through Bayes theorem
- loss and Bayes risk
- Bayes estimate

The basic ingredients - variables

Predict y from measurement x

- x multivariate input
- y multivariate output
- z multivariate hidden/latent variables

introduction of hidden variables often facilitate model specification

The basic ingredients

Problems

- unsupervised modeling if only $\boldsymbol{x}$
- predictive modeling if $\boldsymbol{x}$ and $\boldsymbol{y}$
  - $\boldsymbol{y}$ continuous is regression, e.g., time-series modeling
  - $\boldsymbol{y}$ discrete is classification
- state-space models, mixture models use continuous or discrete hidden variables

The basic ingredients - data

$$\mathcal{D} = \{\boldsymbol{x}_k, \boldsymbol{y}_k\}_{k=1}^{N}$$

Usually i.i.d. samples

The basic ingredients - models

Models

$$p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta},m) = \int p(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta},m)\,d\boldsymbol{z}$$

- $\boldsymbol{\theta}$ are model parameters usually not amenable for interpretation
- m index a particular model structure
- we consider usually flexible universal approximation model families neural networks, Gaussian processes, mixture models

Example

$$p(y|\boldsymbol{x},\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-[y-f(\boldsymbol{x},\boldsymbol{w})]^2/2\sigma^2)$$

$\boldsymbol{\theta} = (\sigma^2, \boldsymbol{w})$

The basic ingredients - priors

$$p(\boldsymbol{\theta})$$

expresses the degree of belief

- probability is limit of frequency #outcomes/#total
- properties beliefs lead to same rules as for probabilities, hence using probability to measure belief

.... more on choice of prior later

The basic ingredients - Bayes theorem

Combining information using Bayes theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, P(B) = \sum_A P(B|A)P(A)$$

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{prob of data}}$$

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{N} p(\boldsymbol{y}_k|\boldsymbol{x}_k, \boldsymbol{\theta}, m)$$

The basic ingredients - predictive distribution

$$p(\boldsymbol{y}|\boldsymbol{x},\mathcal{D},m) = \int p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta},m)\cdot p(\boldsymbol{\theta}|\mathcal{D})\,d\boldsymbol{\theta}$$

is the result of Bayesian learning and provides a full conditional distribution for new inputs $\boldsymbol{x}$

Relation to classical learning

- MAP: $p(\boldsymbol{\theta}|\mathcal{D}) = \delta(\boldsymbol{\theta}-\boldsymbol{\theta}_{MAP})$, $\boldsymbol{\theta}_{MAP} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathcal{D})$

$$p(\boldsymbol{y}|\boldsymbol{x},\mathcal{D},m) = p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}_{MAP},m)$$

- ML: no prior $\boldsymbol{\theta}_{ML} = \arg\max_{\boldsymbol{\theta}} p(\mathcal{D}|\boldsymbol{\theta})$

$$p(\boldsymbol{y}|\boldsymbol{x},\mathcal{D},m) = p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}_{ML},m)$$

The basic ingredients - transductive learning

$$p(\boldsymbol{y}|\boldsymbol{x}, \mathcal{D}, m) = \int p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}, m) \cdot p(\boldsymbol{\theta}|\mathcal{D}, \boldsymbol{x})\, d\boldsymbol{\theta}$$

Model is updated for every test input $\boldsymbol{x}$

$$p(\boldsymbol{\theta}|\mathcal{D}, \boldsymbol{x}) = \frac{p(\mathcal{D}, \boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D}, \boldsymbol{x})}$$

Requires a model for input distribution as well!

Predictive distribution example

Model

$$p(y|\boldsymbol{x},\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x}^\top\boldsymbol{\theta}, \sigma^2)$$

MAP (or ML)

$$p(y|\boldsymbol{x},\mathcal{D}) = \mathcal{N}(\boldsymbol{x}^\top\boldsymbol{\theta}_{MAP}, \sigma^2)$$

Gaussian prior on $\boldsymbol{\theta}$

$$p(y|\boldsymbol{x},\mathcal{D}) = \mathcal{N}(\boldsymbol{x}^\top\widehat{\boldsymbol{\theta}}, \sigma^2 + \boldsymbol{x}^\top\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{x})$$

The basic ingredients - loss function

$$L(\boldsymbol{y}, \widehat{\boldsymbol{y}}) = L(\boldsymbol{y}(\boldsymbol{x}), \widehat{\boldsymbol{y}}(\boldsymbol{x}|\mathcal{D}))$$

defines how close our estimate $\widehat{\boldsymbol{y}}$ is from the truth $\boldsymbol{y}$.
Can formally be defined through axiomatic utility theory Berger

Examples

square loss for continuous variable

$$L(\boldsymbol{y}, \widehat{\boldsymbol{y}}) = (\boldsymbol{y} - \widehat{\boldsymbol{y}})^2$$

zero-one loss for classification $y, \widehat{y} \in [1; C]$

$$L(y, \widehat{y}) = \begin{cases} 0, \ y = \widehat{y} \\ 1, \ y \neq \widehat{y} \end{cases}$$

The basic ingredients - risk

Frequentist risk

$$R(\boldsymbol{y}, \widehat{\boldsymbol{y}}) = \int L(\boldsymbol{y}, \widehat{\boldsymbol{y}}(\mathcal{D})) \cdot p(\mathcal{D})\, d\mathcal{D}$$

average over all possible data sets

Bayes risk

$$\rho(\widehat{\boldsymbol{y}}|\mathcal{D}) = \int L(\boldsymbol{y}, \widehat{\boldsymbol{y}}(\mathcal{D})) \cdot p(\boldsymbol{y}|\boldsymbol{x}, \mathcal{D})\, d\boldsymbol{y}$$

average w.r.t. predictive distribution and conditioned on data

The basic ingredients - risk

Integrated frequency risk

$$
\begin{aligned}
r(\widehat{\boldsymbol{y}}) &= \int L(\boldsymbol{y}, \widehat{\boldsymbol{y}}(\mathcal{D})) \cdot p(\boldsymbol{y}|\boldsymbol{x}, \mathcal{D}) p(\mathcal{D})\, d\boldsymbol{y} d\mathcal{D} \\
&= \int \rho(\widehat{\boldsymbol{y}}|\mathcal{D}) \cdot p(\mathcal{D})\, d\mathcal{D} \\
&= \int L(\boldsymbol{y}, \widehat{\boldsymbol{y}}(\mathcal{D})) \cdot p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta} d\boldsymbol{y}
\end{aligned}
$$

Outline

- Why Bayesian learning?
- Basic ingredients
- Bayes estimators
- More on selection of priors
- Generalization and bias/variance
- Generalization estimation
- Bayesian model selection
- Discussion of Bayesian framework
- Example of Bayesian learning: RVM
- Bayesian signal detection

Bayes estimator

$$\widehat{\boldsymbol{y}}_B = \arg\min_{\widehat{\boldsymbol{y}}} r(\widehat{\boldsymbol{y}}) = \arg\min_{\widehat{\boldsymbol{y}}} \rho(\widehat{\boldsymbol{y}}|\mathcal{D})$$

Risk of Bayes estimator

$$r(\widehat{\boldsymbol{y}}_B)$$

Bayes estimator in continuous case

Square loss

$$\widehat{\boldsymbol{y}}_B = \arg\min_{\widehat{\boldsymbol{y}}} \rho(\widehat{\boldsymbol{y}}|\mathcal{D}) = \arg\min_{\widehat{\boldsymbol{y}}} \int (\boldsymbol{y} - \widehat{\boldsymbol{y}})^2 \cdot p(\boldsymbol{y}|\boldsymbol{x}, \mathcal{D})\, d\boldsymbol{y}$$

$$\widehat{\boldsymbol{y}}_B = \int \boldsymbol{y} \cdot p(\boldsymbol{y}|\boldsymbol{x}, \mathcal{D})\, d\boldsymbol{y} = E_{pred}[\boldsymbol{y}]$$

Absolute loss

$$\widehat{\boldsymbol{y}}_B = \arg\min_{\widehat{\boldsymbol{y}}} \int |\boldsymbol{y} - \widehat{\boldsymbol{y}}| \cdot p(\boldsymbol{y}|\boldsymbol{x}, \mathcal{D})\, d\boldsymbol{y}$$

$\widehat{\boldsymbol{y}}_B$ is the median of predictive distribution

Bayes estimator for classificstion

Loss matrix

Penalty of estimating class $\widehat{y} \in [1; C]$ if the truth is class $y \in [1; C]$

$$L(y, \widehat{y}) = \begin{cases} 0 & if\ \widehat{y} = y \text{ (correct decision)} \\ l(y, \widehat{y}) & if\ \widehat{y} \neq y \in [1; C] \\ t & if\ \widehat{y} = C + 1 \text{ (rejection)} \end{cases}$$

Zero-one loss with rejection

$$L(y, \widehat{y}) = \begin{cases} 0 & if\ \widehat{y} = y \text{ (correct decision)} \\ 1 & if\ \widehat{y} \neq y \in [1; C] \\ t & if\ \widehat{y} = C + 1 \text{ (rejection)} \end{cases}$$

Bayes decision rule - general loss

$$\rho(\widehat{\boldsymbol{y}}|\mathcal{D}) = \sum_y L(y,\widehat{y}) \cdot p(y|\boldsymbol{x},\mathcal{D})$$

For $\widehat{y} = C+1$ then $\sum_y L(y,\widehat{y}) \cdot p(y|\boldsymbol{x},\mathcal{D}) = t \sum_y p(y|\boldsymbol{x},\mathcal{D}) = t$

$$\widehat{y}_B = \begin{cases} k & \text{if } \min_{\widehat{y}\leq C} \sum_y L(y,\widehat{y}) \cdot p(y|\boldsymbol{x},\mathcal{D}) < t \\ C+1 & \text{otherwise} \end{cases}$$

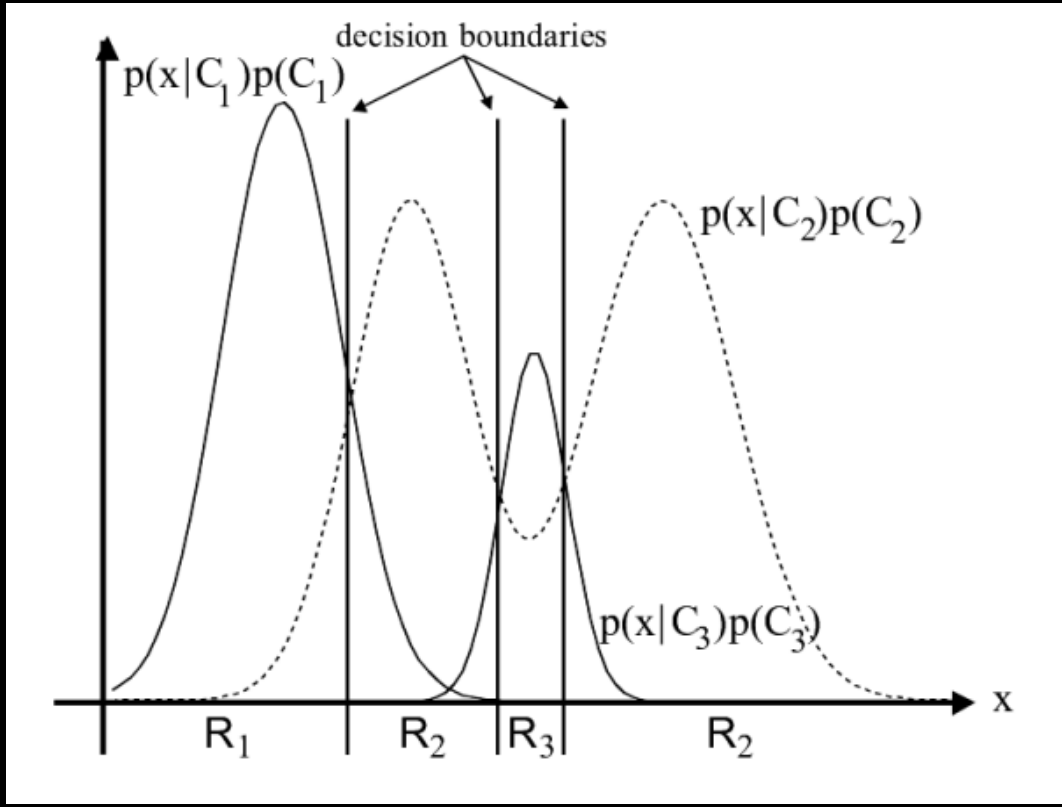
Bayes decision rule - zero-one loss

$$\sum_y L(y,\widehat{y}) \cdot p(y|\boldsymbol{x},\mathcal{D}) = \begin{cases} 1 - p(j|\boldsymbol{x},\mathcal{D}), & \widehat{y} = j \in [1;C] \\ t, & \widehat{y} = C+1 \end{cases}$$

$$\widehat{y}_B = \begin{cases} \arg\max_{y \leq C} p(y|\boldsymbol{x},\mathcal{D}), \text{and prob} > 1-t \\ C+1 \text{ if all } p(y|\boldsymbol{x},\mathcal{D}) \leq 1-t \end{cases}$$

$p(y|\boldsymbol{x},\mathcal{D}) \geq 1/C$ which means $1-t > 1/C$ for rejection to occur

Bayes classifier



decision boundaries are specified by $p(y = i|\boldsymbol{x}, \mathcal{D}) = p(y = j|\boldsymbol{x}, \mathcal{D})$

Optimality through admissibility

Admissibility

$\widehat{\boldsymbol{y}}_0$ is inadmissible if

$$\forall \boldsymbol{y},\ R(\boldsymbol{y}, \widehat{\boldsymbol{y}}_0) \geq R(\boldsymbol{y}, \widehat{\boldsymbol{y}}_1),\ \exists \boldsymbol{y}_0,\ R(\boldsymbol{y}_0, \widehat{\boldsymbol{y}}_0) > R(\boldsymbol{y}_0, \widehat{\boldsymbol{y}}_1)$$

Generalized Bayes estimator is admissible under regularity conditions

- $p(\boldsymbol{y}|\boldsymbol{x}) = p(\boldsymbol{y}|\boldsymbol{x}, \mathcal{D})p(\mathcal{D})/p(\mathcal{D}|\boldsymbol{y}, \boldsymbol{x}) > 0$ for all data
- Bayes risk is finite (might fail for generalized - improper prior case)
- $R(\boldsymbol{y}, \widehat{\boldsymbol{y}})$ is continuous in $\boldsymbol{y}$

Optimality through generalization error

Randomized estimators

use predictive distribution $\widehat{p} = p(\boldsymbol{y}|\boldsymbol{x}, \mathcal{D})$ as random estimator rather than point estimate $\widehat{\boldsymbol{y}}$

Kullback-Leibler information as average loss between distributions

$$\begin{aligned} KL(p|\widehat{p}) &= \int p(\boldsymbol{y}|\boldsymbol{x}) \log \frac{p(\boldsymbol{y}|\boldsymbol{x})}{p(\boldsymbol{y}|\boldsymbol{x}, \mathcal{D})} d\boldsymbol{y} \\ &= E\{L(p, \widehat{p})\} \end{aligned}$$

- the loss is defined as

$$L(p, \widehat{p}) = \log p(\boldsymbol{y}|\boldsymbol{x}) - \log p(\boldsymbol{y}|\boldsymbol{x}, \mathcal{D})$$

- $KL \geq 0$ with 0 if and only if $p \equiv \widehat{p}$

On the property of KL

Inequality

$$-\log\lambda \geq 1-\lambda \text{ for } \lambda > 0 \quad \text{and} \quad -\log\lambda = 1-\lambda \text{ for } \lambda = 1$$

Proof Define $\lambda(y) = \widehat{p}(y)/p(y)$. That is $-\log\lambda(y) = \log p(y)/\widehat{p}(y)$

$$\begin{aligned} KL &= \int p(y)[-\log\lambda(y)]\,dy \\ &\geq \int p(y)[1-\lambda(y)]\,dy \\ &= \int p(y)\left[1-\frac{\widehat{p}(y)}{p(y)}\right]dy \\ &= 0 \end{aligned}$$

Optimality through generalization error

$$KL = \underbrace{G}_{\text{generalization error}} - \underbrace{H(p)}_{\text{entropy}}$$

Average generalization error

$$\Gamma = E_{\mathcal{D}} E_{\boldsymbol{x}}\{G\} = -\int \log p(\boldsymbol{y}|\boldsymbol{x},\mathcal{D}) p(\mathcal{D}) p(\boldsymbol{x})\, d\boldsymbol{y} d\boldsymbol{x} d\mathcal{D}$$

is integrated frequency risk (also averaged w.r.t $p(\boldsymbol{x})$) up to a constant

L.K. Hansen: "Bayesian Averaging is Well-Temperated," NIPS99, 265–271, 2000 shows optimality in generalization error

Outline

- Why Bayesian learning?
- Basic ingredients
- Bayes estimators
- More on selection of priors
- Generalization and bias/variance
- Generalization estimation
- Bayesian model selection
- Discussion of Bayesian framework

More on priors

- **subjective priors**: consider relative likelihood of various parameters values
- **empirical priors**: obtained from past experience data
- **structural priors**:
  - independence of some parameters?
  - imposing functional smoothness
  - invoking constraints
- **convenience priors**:
  - nice functional form in order to make calculations simple
  - conjugate priors: posterior and prior have same shape exponential family is important

More on priors

- **hierarchical**: $p(\theta) = \int p(\theta|\lambda)p(\lambda)\,d\lambda$
- **non-informative**: make the influence of the prior as small as possible
- **improper**: improper priors do not integrate to one. Leads to generalized Bayes estimator which typically also is admissible

Non-informative priors

- discrete parameter taking C values: $p(\theta) = 1/C$
- continuous parameter $p(\theta) = 1$ which is improper $\int p(\theta)\, d\theta = \infty$

Location parameter

invariance to choice of parameterization

$\eta = \theta + c, \forall c$

$$\int p(\eta)\, d\eta = \int p(\theta)\, d\theta$$

$$\int p(\eta - c)\, d\eta = \int p(\eta)\, d\eta$$

for all η. With $\eta = c$ then $p(c) = p(0)$ thus $p(\theta) = 1$

Non-informative priors

Scale parameter $\eta = c\theta, \forall c > 0$

$$\int p(\eta)\, d\eta = \int p(\theta)\, d\theta$$

$$\int p(\eta c^{-1}) c^{-1}\, d\eta = \int p(\eta)\, d\eta$$

thus with $\eta = c$ and $p(c) = c^{-1}p(1)$. Setting $p(1) = 1$ then $p(\theta) = \theta^{-1}$

Jeffrey's non-informative

$$p(\boldsymbol{\theta}) = \sqrt{\det \boldsymbol{I}(\boldsymbol{\theta})}$$

$$\boldsymbol{I}(\boldsymbol{\theta}) = -E[\partial^2 \log p(\boldsymbol{x}|\theta)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top]$$

Optimizing hyperparaters using evidence

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{\lambda})$$

$\boldsymbol{\lambda}$ are **hyperparameters**

Evidence - marginal likelihood

$$p(\mathcal{D}) = p(\mathcal{D}|\boldsymbol{\lambda}) = \int p(\mathcal{D}|\boldsymbol{\theta}, \boldsymbol{\lambda}) p(\boldsymbol{\theta}|\boldsymbol{\lambda}) \, d\boldsymbol{\theta}$$

$$\boldsymbol{\lambda}_{ML-II} = \arg\max_{\boldsymbol{\lambda}} p(\mathcal{D}|\boldsymbol{\lambda})$$

Outline

- Why Bayesian learning?
- Basic ingredients
- Bayes estimators
- More on selection of priors
- Generalization and bias/variance
- Generalization estimation
- Bayesian model selection
- Discussion of Bayesian framework
- Example of Bayesian learning: RVM
- Bayesian signal detection

Generalization is the ultimate frequentist objective

Assume that the underlying system is stationary (time-invariant).
How well are we doing on future data?

Generalization error

$$G(\mathcal{D}) = \int L(\boldsymbol{y}, \widehat{\boldsymbol{y}}(\mathcal{D})) \cdot p(\boldsymbol{x}, \boldsymbol{y})\, d\boldsymbol{x} d\boldsymbol{y}$$

- $\{\boldsymbol{x}; \boldsymbol{y}\}$ is a sample independent of all samples in the training set
- $L(\cdot)$ is any loss function

Generalization error decomposition

Average generalization error / averaged integrated risk

$$\Gamma = E_{\mathcal{D}}\{G(\mathcal{D})\}$$

Decomposition

$$\Gamma = \textit{Inherent Noise} + \textit{Bias} + \textit{Variance}$$

- *Inherent Noise* (minimal Bayes risk) can not be modeled
- *Bias* is due to an incomplete model
- *Variance* is due to a finite training set

A general B/V decomposition Heskes, 1998

Properties

- bias only depends on true and average distributions
- variance is non-negative not a function of true distribution, and zero only if and only if distributions are equal
- mean-square error is a special case

Decomposition

$$\begin{aligned}\Gamma = E_{\mathcal{D}}\left\{G(\mathcal{D})\right\} &= H(p) + KL(p|\widehat{p}) \\ &= H(p) + KL(p|\bar{p}) + E_{\mathcal{D}}\left\{KL(\bar{p}|\widehat{p})\right\} \\ &= \textit{Inherent Noise} + \textit{Bias} + \textit{Variance}\end{aligned}$$

with average model

$$\bar{p} = Z^{-1} \cdot \exp(E_{\mathcal{D}}\left\{\log \widehat{p}\right\})$$

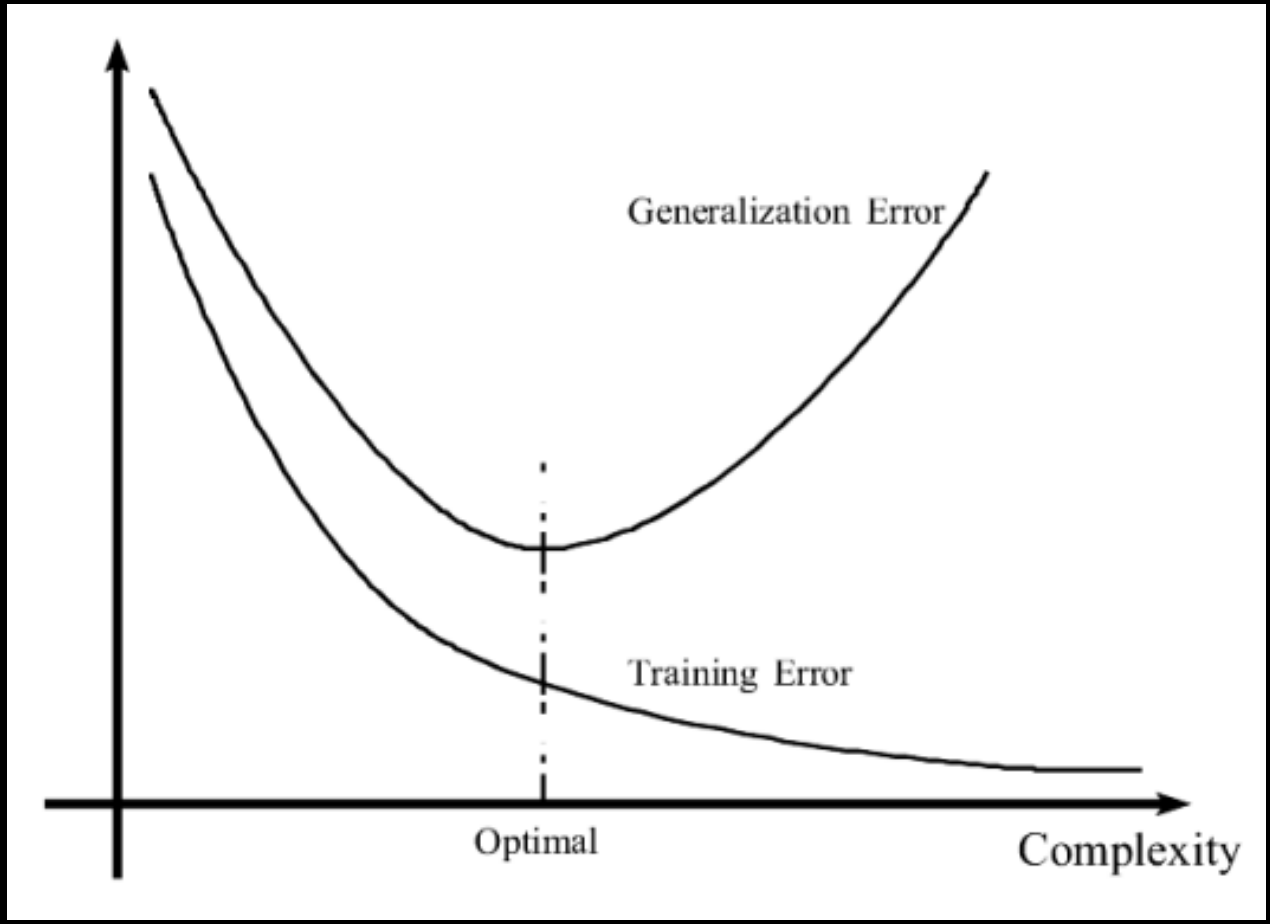
Bias/variance dilemma

The model should learn rather than memorize training data

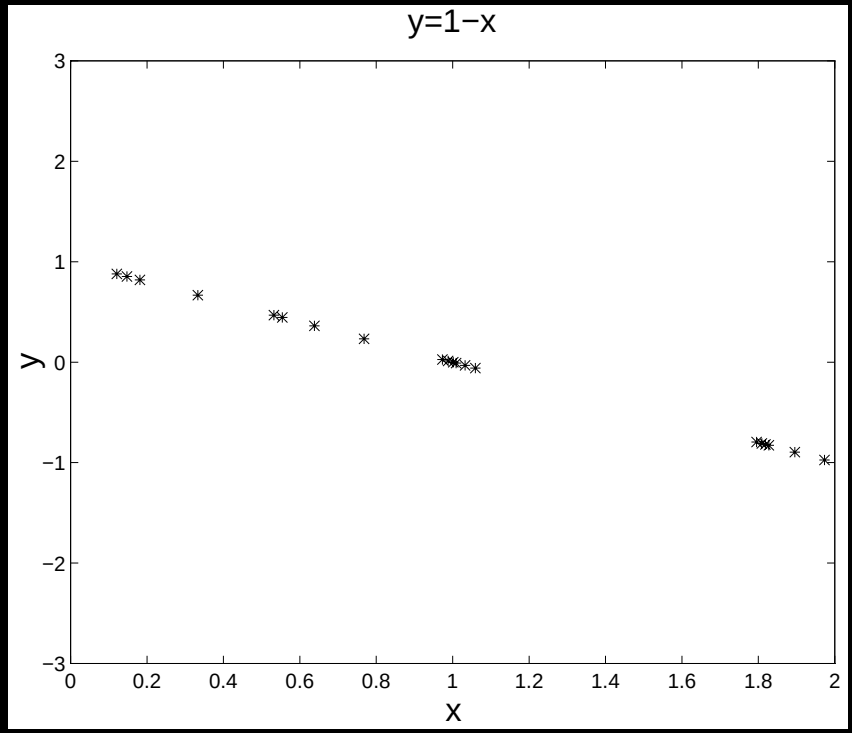
Model complexity

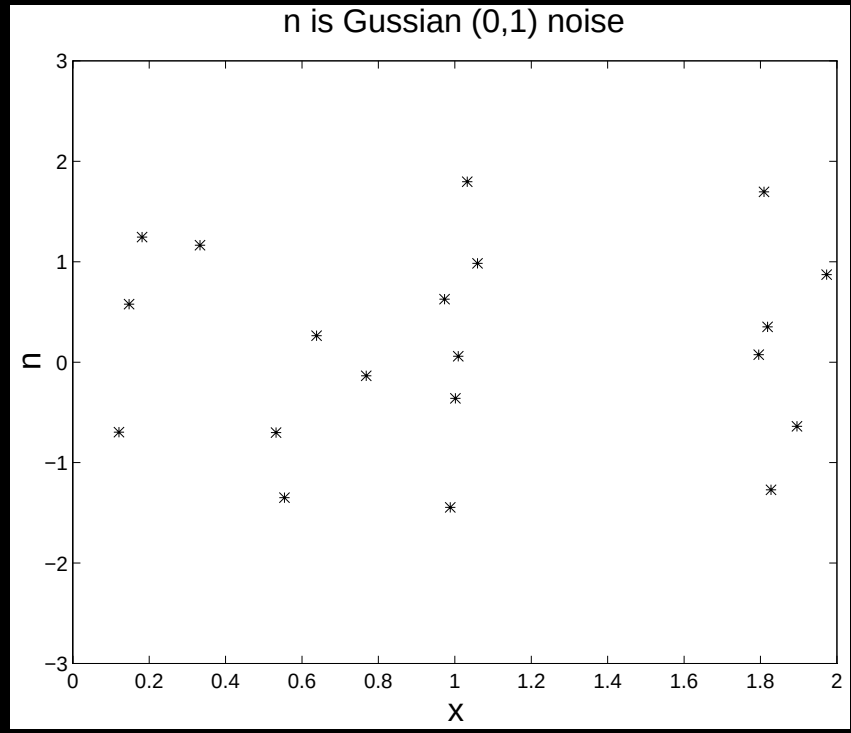
low	high
cannot fit	fits to noise

Bias/variance dilemma

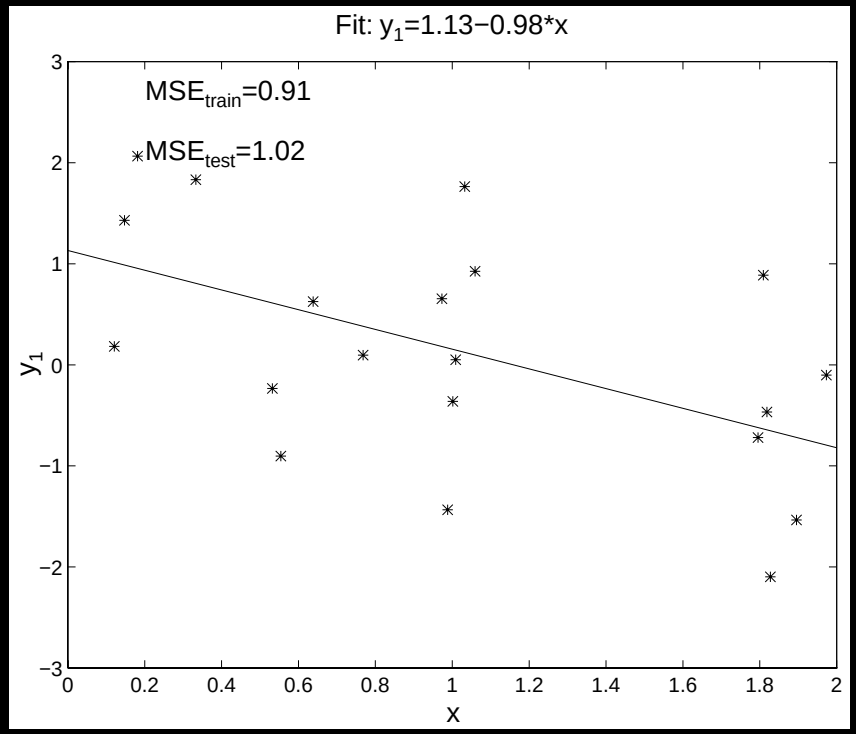


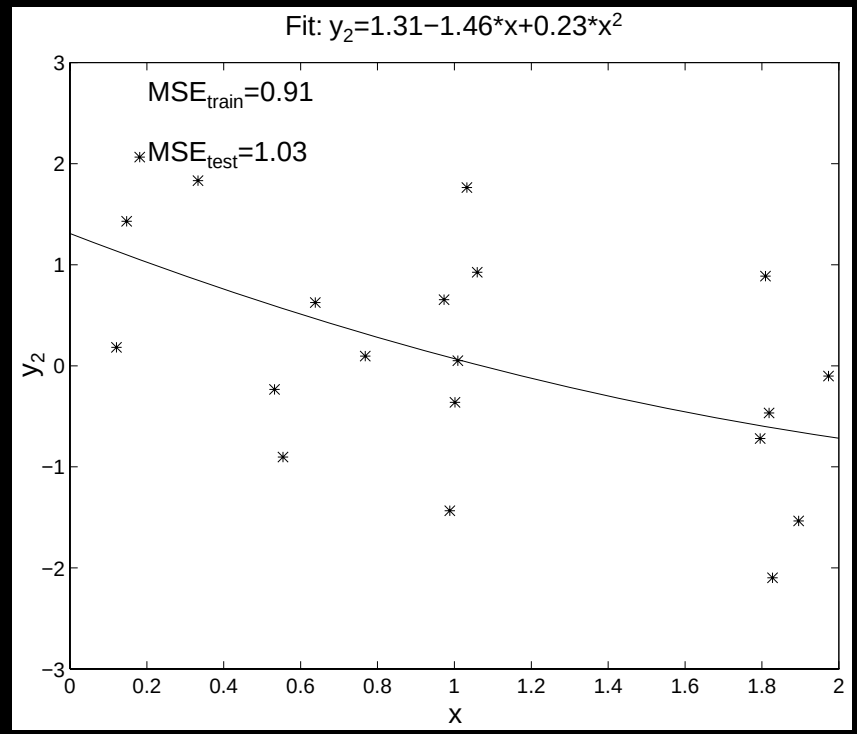
Overfitting



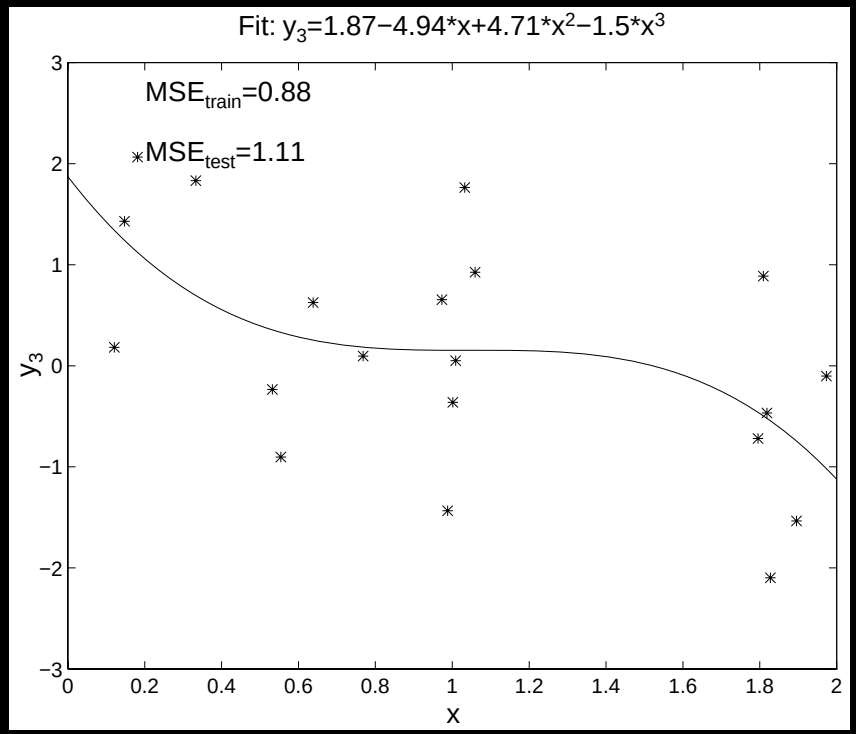


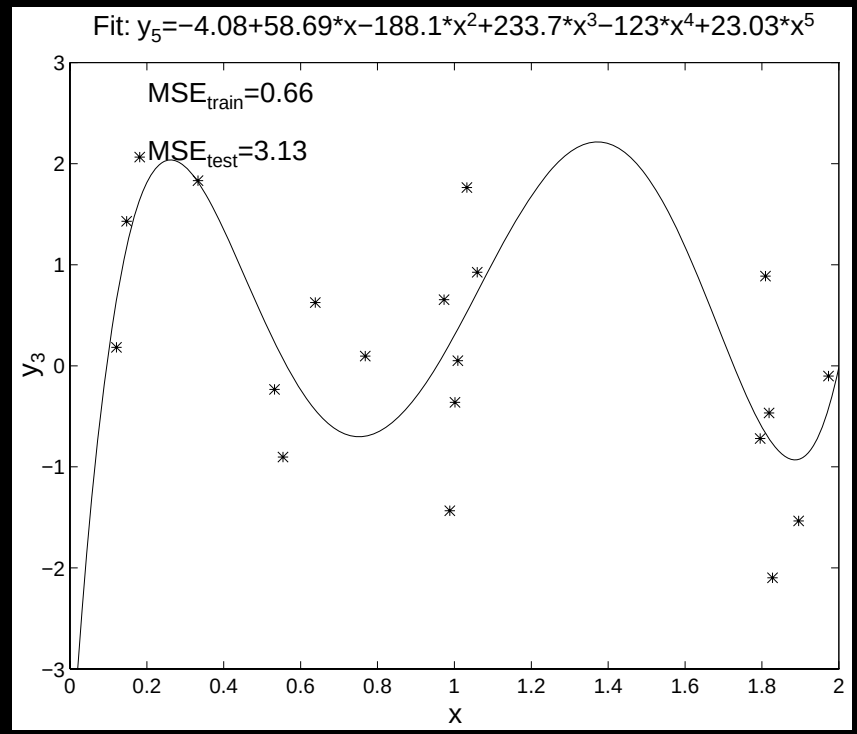
Overfitting



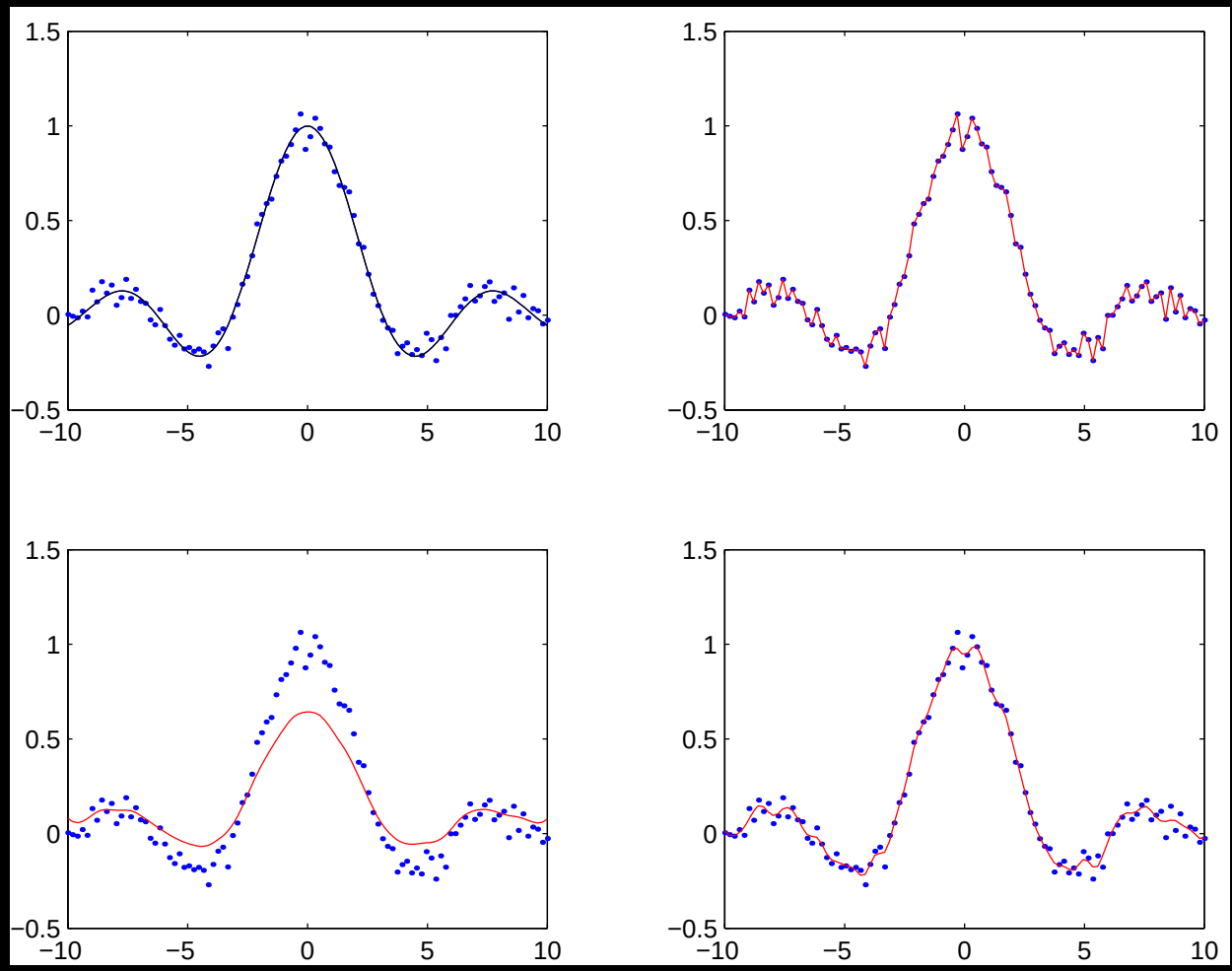


Overfitting





Overfitting in an RBF network



Outline

- Why Bayesian learning?
- Basic ingredients
- Bayes estimators
- More on selection of priors
- Generalization and bias/variance
- Generalization estimation
- Bayesian model selection
- Discussion of Bayesian framework
- Example of Bayesian learning: RVM
- Bayesian signal detection

Generalization estimation

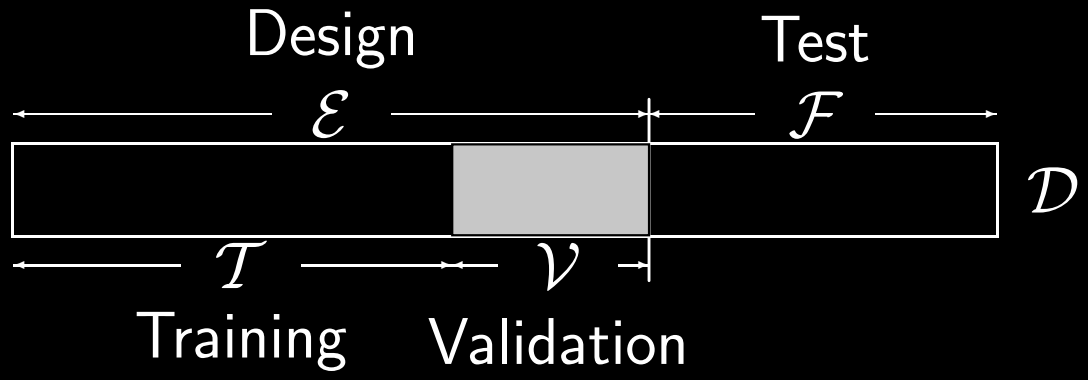
Approaches

- Asymptotic theory leading to algebraic estimates of the average generalization error
- Resampling approaches (cross-validation, jackknife, and bootstrap) of the generalization error or average generalization error

Purposes

- Assessing the final quality and reliability of the model
- Model selection

Limited data is always a challenge

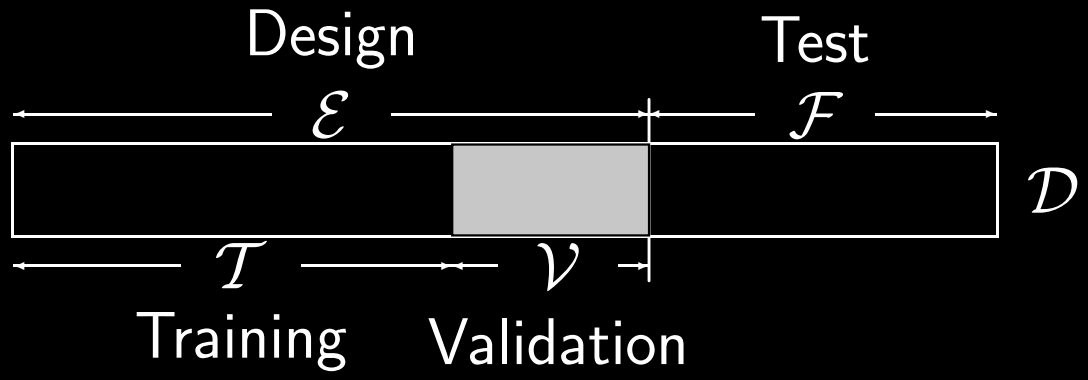


Design/Test Split

- Test set is exclusively used for final assessment of model designed from $\mathcal{E}$
- Objective is high generalization ability and reliable assessment

J. Larsen and C. Goutte: "On Optimal Data Split for Generalization Estimation and Model Selection," in Proceedings of the IEEE NNSP Workshop IX, pp. 225–234, 1999

Limited data is always a challenge



Training/Validation Split

- Model is trained on training set. Validation set is used to select optimal model or tune additional hyperparameters
- Objective is high generalization ability.

Hold-Out Cross-Validation

Data Splitting

- $\gamma = i/N \; i = 1, 2, \cdots, N-1$ is the *split ratio*
- $N_{\mathcal{F}} = \gamma N$ for testing and $N_{\mathcal{E}} = (1-\gamma)N$ for design

HO Estimate

$$\widehat{G}_{\mathrm{HO}} = {N_{\mathcal{F}}}^{-1} \sum_{k \in \mathcal{F}} -\log p(\boldsymbol{y}(k)|x(k), \mathcal{E}, m)$$

Property HO is an unbiased estimate of the generalization error for i.i.d. samples

Quality of the HO Estimator

$$MSE_{HO}(\gamma) = E_{\mathcal{D}}\left\{\left(\widehat{G}_{\mathrm{HO}} - G^*\right)^2\right\}$$

$$= \underbrace{E_{\mathcal{D}}\left\{\left(\widehat{G}_{HO} - G(\mathcal{D})\right)^2\right\}}_{variance} + \underbrace{E_{\mathcal{D}}\left\{(G(\mathcal{D}) - G^*)^2\right\}}_{bias}$$

where G^* is the minimum achievable gen. error for the current model, i.e., infinitely data

Property

- Bias $\downarrow$ as $\gamma \downarrow$
- Variance $\uparrow$ as $\gamma \downarrow$

K-fold Cross-Validation

Procedure

- Split data into K disjoint subsets $\mathcal{F}_j$ (approx. equal sizes)
- $K = \lfloor 1/\gamma \rfloor$, for $\gamma < 1/2$,
  $\gamma = \{1/N, 1/(N-1), \cdots, 1/2, \cdots, 1 - 1/N\}$
- Evaluate on each subset the model designed on the remaining data, $\mathcal{E}_j = \mathcal{D} \setminus \mathcal{F}_j$

Estimate

$$\widehat{\Gamma}_{\mathrm{KCV}} = \frac{1}{N} \sum_{j=1}^{K} \sum_{k \in \mathcal{F}_j} -\log p(\boldsymbol{y}(k)|x(k), \mathcal{E}_j, m)$$

Property Unbiased estimator the average generalization error, Γ, based on $N_{\mathcal{E}}$ data.

Quality of KCV Estimator

$$MSE_{KCV}(\gamma) = E_{\mathcal{D}}\left\{\left(\widehat{\Gamma}_{\mathrm{KCV}} - G^*\right)^2\right\}$$

$$= \underbrace{E_{\mathcal{D}}\left\{\left(\widehat{\Gamma}_{\mathrm{KCV}} - \Gamma\right)^2\right\}}_{\text{variance}} + \underbrace{E_{\mathcal{D}}\left\{(\Gamma - G^*)^2\right\}}_{\text{bias}}$$

Results

- Analytical expression for opt. design/test splits for location parameter model. Tends to hold asymp. for other models:
  - HO: $\gamma_{opt} \rightarrow 1$, as $N \rightarrow \infty$
  - KCV: $\gamma_{opt} = 1/N$, (LOO)
- Model selection using KCV:
  - $\gamma_{opt} \rightarrow 1$, as $N \rightarrow \infty$
  - LOO seems to optimal when N is small both wrt. generalization error and probability of selecting correct model
- Model selection using HO:
  - Conflict between opt. gen. error and probability of selecting correct model

Algebraic Generalization Error Approach

Properties

- Asymptotic estimates valid for large training sets
- Various assumptions on model bias and example dependencies
- No data need to be set aside for validation

Estimator	Model	Est.
Exact (Hansen 93)	Unbiased lin. zero-mean Gaus data	ML
FPE (Akaike 69)	Unbiased, no prior	MSE
FPER (Larsen 94)	Unbiased with prior	Pen. MSE
AIC (Akaike 73)	Unbiased no prior	ML
AICc (Hurvich&Tsai 1989)	Unbiased no prior	ML
GEN (Larsen 92, 2000)	No restrictions, auto corr data	Pen. MSE/MAP
GPE (Moody 91)	nonlin with prior	MSE
NIC (Murata 94)	nonlin(NN),nested, i.i.d. data	MAP
TIC (Takeuchi 76)	nonlin i.i.d. data	MAP
GIC (Konishi&Kitagawa 96)	general i.i.d.	MAP,Bayes
DIC (Spiegelhalter et al. 2002)	general i.i.d.	Bayes

GEN

Major Assumptions

- Asymptotic validity, $o(1/N)$
- Applies to bias and regularized models
- Estimates average generalization error

MAP approach

$$C_{\mathcal{D}}(\boldsymbol{\theta}) = N^{-1} \sum_{k=1}^{N} \ell(\boldsymbol{y}(k)|\boldsymbol{x}(k), \boldsymbol{\theta}) + R(\boldsymbol{\theta}) = S_{\mathcal{D}}(\boldsymbol{\theta}) + R(\boldsymbol{\theta})$$

GEN Estimator

$$\Gamma_{GEN} = E_{\mathcal{D}}\{S_{\mathcal{D}}(\widehat{\boldsymbol{\theta}})\} + \frac{m_{eff}}{N} - \frac{A}{N} \cdot \frac{\partial R}{\partial \boldsymbol{\theta}^{\top}}(\boldsymbol{\theta}^*)\boldsymbol{J}^{-1}(\boldsymbol{\theta}^*)\frac{\partial R}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}^*)$$

- Optimal parameters: $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} G(\boldsymbol{\theta})$ where the expected cost: $C(\boldsymbol{\theta}) = E_{\mathcal{D}}\{C_{\mathcal{D}}(\boldsymbol{\theta})\} = G(\boldsymbol{\theta}) + R(\boldsymbol{\theta})$
- For practical use an unbiased $o(1/N)$ estimator is obtained by, neglecting the expectation, replacing $\boldsymbol{\theta}^*$ by $\widehat{\boldsymbol{\theta}}$, and $\boldsymbol{J}$ by $\boldsymbol{J}_{\mathcal{D}}$

GEN Estimator

- Effective number of parameters

$$m_{eff} = \text{tr}\left[\boldsymbol{J}^{-1}(\boldsymbol{\theta}^*)\left(\boldsymbol{K}(0) + \sum_{n=1}^{\bar{M}} \frac{N-n}{N}(\boldsymbol{K}(n) + \boldsymbol{K}^\top(n))\right)\right]$$
$$= \text{tr}\left[\boldsymbol{J}^{-1}(\boldsymbol{\theta}^*)\boldsymbol{L}\right]$$

  where $\bar{M} = \min(M, N-1)$, M is the time dependence length (for i.i.d. examples $M = 0$),
  $A = \bar{M} + 1 - \bar{M}(\bar{M}+1)/2N$

- $\boldsymbol{K}(n) = E\{\partial\ell(k)/\partial\boldsymbol{\theta} \cdot \partial\ell(k+n)/\partial\boldsymbol{\theta}^\top\}$ with $\ell(k) \equiv \ell(\boldsymbol{y}(k)|\boldsymbol{x}(k), \boldsymbol{\theta}^*))$
- $\boldsymbol{J}(\boldsymbol{\theta})$ is the Hessian matrix of the expected cost function $C(\boldsymbol{\theta})$

Outline

- Why Bayesian learning?
- Basic ingredients
- Bayes estimators
- More on selection of priors
- Generalization and bias/variance
- Generalization estimation
- Bayesian model selection
- Discussion of Bayesian framework
- Example of Bayesian learning: RVM
- Bayesian signal detection

Bayesian model selection

Bayes optimal decision rule (under 0/1 loss function) leads to the optimal model

$$m_{opt} = \arg\max_m p(m|\mathcal{D})$$

$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)P(m)}{\sum_{m=1}^{M} p(\mathcal{D}|m)P(m)}$$

is the probability of the model given data

The Bayes model probability is only correct if the true model is among the hypothesis models!

Probabilistic Model Selection

Uniform Model Belief

In the case of equal model priors, i.e., $P(m) = 1/M$, the model selection concerns computing the **evidence** $p(\mathcal{D}|m)$

Evidence

$$p(\mathcal{D}|m) = \int p(\mathcal{D}, \boldsymbol{\theta}|m)\, d\boldsymbol{\theta} = \int p(\mathcal{D}|\boldsymbol{\theta}, m) p(\boldsymbol{\theta}|m)\, d\boldsymbol{\theta}$$

where

- $\boldsymbol{\theta}$ are model parameters
- $p(\mathcal{D}|\boldsymbol{\theta}, m)$ is the likelihood
- $p(\boldsymbol{\theta}|m)$ is the prior which is normally assumed vague and normalizable

Approximations

- Laplace approximation
- BIC approximation (large sample Laplace)
- variational Bayes ("EM like")
- expectation propagation
- ensemble learning
- (annealed) importance sampling, particle filtering
- Gibbs sampling
- Markov chain Monte Carlo methods (Metropolis-Hastings, Parallel tempering (gets marginal likelihood), particle path filter)

Models with hidden variables

$$p(\boldsymbol{y}|\boldsymbol{x},\mathcal{D}) = \int p(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta},m)\cdot p(\boldsymbol{\theta}|\mathcal{D})\,d\boldsymbol{z}d\boldsymbol{\theta}$$

In particular ensemble learning and Variational Bayes are useful

- approximate by integrating w.r.t. proposal distributions $q(\boldsymbol{z},\boldsymbol{\theta})$
- ensemble learning uses simple functional forms often fully factorized $q(\boldsymbol{z},\boldsymbol{\theta}) = \prod_i q(z_i)\prod_j q(\theta_j)$
- Variational Bayes uses functional forms as priors partly factorized $q(\boldsymbol{z},\boldsymbol{\theta}) = q(\boldsymbol{z})\prod_j q(\theta_j)$
- works for on-line models Ghahramani, Valpola

Variational Bayes learning

Key ingredients

- Use factorized approximate posterior distribution

$$q(\boldsymbol{\theta}, \boldsymbol{z}) = q(\boldsymbol{\theta}) \cdot q(\boldsymbol{z})$$

- Use Jensen's inequality to bound marginal likelihood (evidence)

Resources

- www.variational-bayes.org
- Z. Ghahramani, C. Bishop, G. Hinton, M.I. Jordan, D. MacKay, C. Rasmussen, R. Neal, M. Beal

Variational Bayes learning

$$\begin{aligned}\log p(\mathcal{D}|m) &= \log \int p(\mathcal{D}, \boldsymbol{\theta}, \boldsymbol{z}|m)\, d\boldsymbol{z} d\boldsymbol{\theta} \\ &= \log \int \frac{q(\boldsymbol{\theta})q(\boldsymbol{z})p(\mathcal{D}, \boldsymbol{\theta}, \boldsymbol{z}|m)}{q(\boldsymbol{\theta})q(\boldsymbol{z})}\, d\boldsymbol{z} d\boldsymbol{\theta} \\ &= \mathcal{F}_m(q(\boldsymbol{\theta}), q(\boldsymbol{z}), \mathcal{D}) + KL_{post}(q\|p)\end{aligned}$$

Variational free energy $\mathcal{F}$ bounds the evidence

$$\begin{aligned}\log p(\mathcal{D}|m) &\geq \mathcal{F}_m(q(\boldsymbol{\theta}), q(\boldsymbol{z}), \mathcal{D}) \\ &= \int q(\boldsymbol{\theta})q(\boldsymbol{z}) \log \frac{p(\mathcal{D}, \boldsymbol{\theta}, \boldsymbol{z}|m)}{q(\boldsymbol{\theta})q(\boldsymbol{z})}\, d\boldsymbol{z} d\boldsymbol{\theta}\end{aligned}$$

Variational Bayes learning

Kullback-Liebler between true and approximate posterior

$$KL_{post}(q|p) = \int q(\boldsymbol{\theta})q(\boldsymbol{z}) \log \frac{p(\boldsymbol{\theta}, \boldsymbol{z}|\mathcal{D}, m)}{q(\boldsymbol{\theta})q(\boldsymbol{z})} d\boldsymbol{z}d\boldsymbol{\theta}$$

Variational Bayes learning

Maximize $\mathcal{F}_m(q(\boldsymbol{\theta}), q(\boldsymbol{z}), \mathcal{D})$ w.r.t. $q(\boldsymbol{\theta})$ and $q(\boldsymbol{z})$ i.e., minimize $KL_{post}(q\|p)$

"E-step" - estimate posterior over hidden variables

$$q^{(j+1)}(\boldsymbol{z}) \propto \exp\left[\int \log p(\mathcal{D}, \boldsymbol{z}|\boldsymbol{\theta}, m) q^{(j)}(\boldsymbol{\theta})\right] d\boldsymbol{\theta}$$

"M-step" - estimate posterior over parameters

$$q^{(j+1)}(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}) \exp\left[\int \log p(\mathcal{D}, \boldsymbol{z}|\boldsymbol{\theta}, m) q^{(j+1)}(\boldsymbol{z})\right] d\boldsymbol{z}$$

Reduces to classical EM when $q(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$

Exponential distribution family

Complete likelihood

$$p(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{x}) = f(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}) g(\boldsymbol{\theta}) \exp\left[\boldsymbol{\phi}(\boldsymbol{\theta})^\top \boldsymbol{u}(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{x})\right]$$

Conjugate prior

$$p(\boldsymbol{\theta}|\eta, \boldsymbol{\nu}) = h(\eta, \boldsymbol{\nu}) g(\boldsymbol{\theta})^\eta \exp\left[\boldsymbol{\phi}(\boldsymbol{\theta})^\top \boldsymbol{\nu}\right]$$

- many standard distribution belongs to exponential family
- also complete likelihood for many mixture models, classes of Markov models, etc.

Optimum under exponential distribution family

i.i.d. Data $\mathcal{D} = \{\boldsymbol{x}_k, \boldsymbol{y}_k\}_{k=1}^N$

Exact posterior as in regular EM but using averaged natural parameters

$$q(\boldsymbol{z}_k) \propto f(\boldsymbol{y}_k, \boldsymbol{x}_k, \boldsymbol{z}_k) \exp\left[\bar{\boldsymbol{\phi}}(\boldsymbol{\theta})^\top \boldsymbol{u}(\boldsymbol{y}_k, \boldsymbol{z}_k, \boldsymbol{x}_k)\right] = p(\boldsymbol{z}_k | \boldsymbol{y}_k, \boldsymbol{x}_k, \bar{\boldsymbol{\phi}}(\boldsymbol{\theta}))$$

$\bar{\boldsymbol{\phi}}(\boldsymbol{\theta}) = \langle \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle_{q(\boldsymbol{\theta})}$ are natural parameters

$q(\boldsymbol{\theta})$ also exponential family conjugate

$$q(\boldsymbol{\theta}) = h(\widetilde{\eta}, \widetilde{\boldsymbol{\nu}}) g(\boldsymbol{\theta})^{\widetilde{\eta}} \exp\left[\boldsymbol{\phi}(\boldsymbol{\theta})^\top \widetilde{\boldsymbol{\nu}}\right]$$

$\widetilde{\eta} = \eta + N$, $\widetilde{\boldsymbol{\nu}} = \boldsymbol{\nu} + \sum_{n=1}^N \bar{\boldsymbol{u}}(\boldsymbol{y}_k, \boldsymbol{x}_k)$,
$\bar{\boldsymbol{u}}(\boldsymbol{y}_k, \boldsymbol{x}_k) = \langle \boldsymbol{u}(\boldsymbol{y}_k, \boldsymbol{x}_k) \rangle_{q(\boldsymbol{z})}$

Expectation propagation

- Minka: focus on approximating maginals for each sample: $t_k(\boldsymbol{\theta}) = p(\boldsymbol{\theta})p(y_k|\boldsymbol{x}_k, \boldsymbol{\theta})$.
- Use $KL(p|q)$ not $KL(q|p)$ as in VB which typically under-estimates variability.
- Iterate for each sample k
  - Deletion: delete $t_k(\boldsymbol{\theta})$
  - Projection: update $\tilde{t}_k(\boldsymbol{\theta})$
  - Inclusion: update $q(\boldsymbol{\theta})$
- No proof of convergence.

Simple evidence approximation

Simpler than more involved methods like Laplace, variational Bayes and MCMC

Normalized log-posterior

$$C_{\mathcal{D}}(\boldsymbol{\theta}) = \frac{1}{N}\left(\log p(\mathcal{D}|\boldsymbol{\theta}, m) + \log p(\boldsymbol{\theta}|m)\right)$$

and the maximum a posteriori (MAP) solution

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} C_{\mathcal{D}}(\boldsymbol{\theta})$$

Approximations

Gaussian MAP approximation

$$C_{\mathcal{D}}(\boldsymbol{\theta}) = C_{\mathcal{D}}(\widehat{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})\boldsymbol{J}_{\mathcal{D}}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}})^{\top}$$

$$\boldsymbol{J}_{\mathcal{D}} = -\frac{\partial^2 C_{\mathcal{D}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\top}}\bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = O(1)$$

Essential assumptions

- $\boldsymbol{J}_{\mathcal{D}}$ should be of full rank, hence $\boldsymbol{J}_{\mathcal{D}}^{-1}$ should exists
- $\boldsymbol{J}_{\mathcal{D}} = O(1)$ is usually fulfilled with N^{-1} normalization. Sinusoidal model is a counter example Stoica

Approximations

Laplace Approximation

$$
\begin{aligned}
p(\mathcal{D}|m) &= \int \exp\left(N C_{\mathcal{D}}(\boldsymbol{\theta})\right) d\boldsymbol{\theta} \\
&\approx \int \exp\left(N C_{\mathcal{D}}(\widehat{\boldsymbol{\theta}}) - \frac{N}{2}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})\boldsymbol{J}_{\mathcal{D}}(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}})^{\top}\right) d\boldsymbol{\theta} \\
&\approx p(\mathcal{D}|\widehat{\boldsymbol{\theta}}, m) \cdot p(\widehat{\boldsymbol{\theta}}|m) \cdot \left(\frac{2\pi}{N}\right)^{\frac{\dim(\boldsymbol{\theta})}{2}} \cdot |\boldsymbol{J}|_{\mathcal{D}}^{-\frac{1}{2}}
\end{aligned}
$$

Bayesian Information Criterion

Since $\boldsymbol{J}_{\mathcal{D}} = O(1)$, the leading term for large N does not involve the often complicated Hessian, hence, the evidence is approximated as

$$p(\mathcal{D}|m) \approx p(\mathcal{D}|\widehat{\boldsymbol{\theta}}, m) \cdot p(\widehat{\boldsymbol{\theta}}|m) \cdot \left(\frac{2\pi}{N}\right)^{\frac{\dim(\boldsymbol{\theta})}{2}}$$

$$\log p(D|m)/N \approx \mathrm{BIC} = C_{\mathcal{D}}(\widehat{\boldsymbol{\theta}}) + \dim(\widehat{\boldsymbol{\theta}}) \cdot \log(N)/(2N)$$

Views on Bayesian model selection

- BIC gives a consistent model selection as $N \rightarrow \infty$ if the true model is among the candidates
- GEN/AIC consistently overfit for $N \rightarrow \infty$ has a smaller penalty $\dim(\boldsymbol{\theta})/N$ compared to $\dim(\boldsymbol{\theta}) \cdot \log(N)/2N$ in BIC

Connection between BIC and GEN/AIC

- In GEN/AIC $E_{\mathcal{D}}\{G(\mathcal{D})\}$ is approximated by a 2nd order Taylor
- In BIC, $E_{\mathcal{D}}\{\exp(G(\mathcal{D}))\}$ is approximated a 2nd order Taylor; hence $\log$ is performed

Outline

- Why Bayesian learning?
- Basic ingredients
- Bayes estimators
- More on selection of priors
- Generalization and bias/variance
- Generalization estimation
- Bayesian model selection
- Discussion of Bayesian framework
- Example of Bayesian learning: RVM
- Bayesian signal detection

Discussion of Bayesian Learning

Objectivity

We want to be as objective as possible, however without prior expectation nothing can be learned

- no-free-lunch theorems
- link to philosophical theories
- J. Friedman: "no methods dominates all others over all possible situations"

Discussion of Bayesian Learning

Prior knowledge

likelihood, loss function, model family, parameter priors

- G.E.P. Box (1976) and Stephen Strother: "all models are wrong – but some are useful"
- use flexible models with careful model optimization
- be as data-driven as possible, minimum non-informative prior assumptions
- use Bayes for formal incorporation of all available knowledge
- use careful model evaluation (generalization performance, robustnes to changes in assumptions, sensitivity analysis)

Discussion of Bayesian Learning

Robustness

Slight changes in model assumptions should lead to slight changes in conclusions/decisions

- sensitivity analysis
- generalization error - test performance
- extensive cross-validation
- learning curves
- information conveyed by data and by prior - if they clash we want likelihood dominance.
- errorbars and fluctuations in predictive distributions

Discussion of Bayesian Learning

Robustness

- iterated modeling until desired performance/robustness is obtained
- trade-off between performance and robustness for specific limited data set

DTU

Outline

- Why Bayesian learning?
- Basic ingredients
- Bayes estimators
- More on selection of priors
- Generalization and bias/variance
- Generalization estimation
- Bayesian model selection
- Discussion of Bayesian framework
- Example of Bayesian learning: RVM
- Bayesian signal detection

Bayesian learning in RBF nets - relevance vector machine

RBF network

$$y = \sum_{j=1}^{n_H} \phi_j(\boldsymbol{x}) w_{ij}^O + \varepsilon = \boldsymbol{\phi}^\top(\boldsymbol{x})\boldsymbol{\theta} + \varepsilon$$

- $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and i.i.d
- $\phi_j(\boldsymbol{x}) = \exp(-\|\boldsymbol{x} - \boldsymbol{x}(j)\|^2/2\nu^2)$,

Vector notation of training data

$$\boldsymbol{y} = \boldsymbol{\Phi}^\top \boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

Getting the predictive distribution

Ingredients

Prior	$p(\boldsymbol{\theta}\|\boldsymbol{A}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{A}^{-1})$, $\boldsymbol{A} = diag(\boldsymbol{\alpha})$, α_j is (inverse) individual weight decay or hyperparameter
Likelihood	$p(\mathcal{D}\|\boldsymbol{\theta}, \sigma^2, \nu^2) = \prod_{k=1}^{N} \mathcal{N}(y(k) - \boldsymbol{\phi}^\top(\boldsymbol{x}(k))\boldsymbol{\theta}, \sigma^2)$

Posterior weight distribution

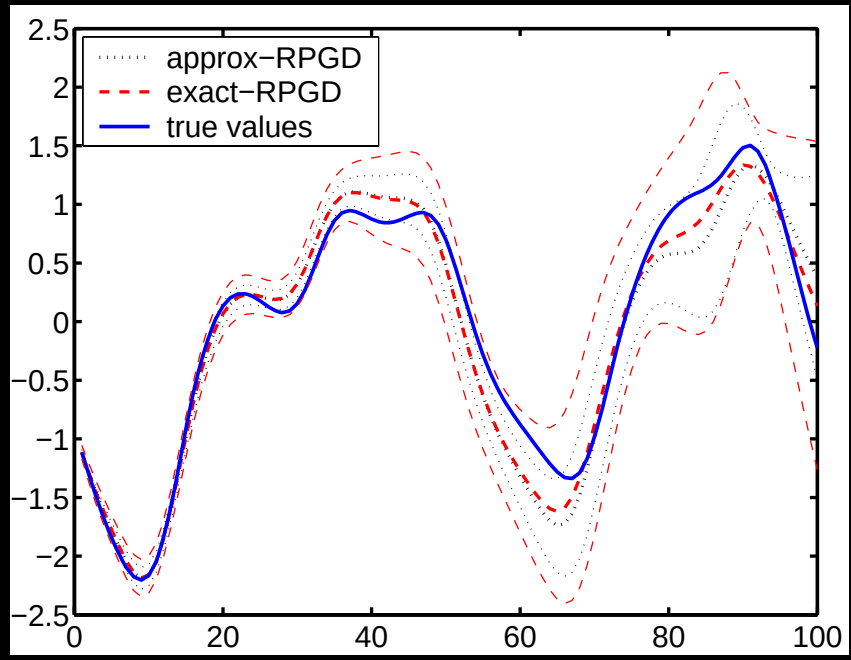
$$p(\boldsymbol{\theta}|\mathcal{D}, \sigma^2, \boldsymbol{A}, \nu^2) = \frac{p(\mathcal{D}|\boldsymbol{\theta}, \sigma^2, \nu^2)p(\boldsymbol{\theta}|\boldsymbol{A})}{p(\mathcal{D}|\sigma^2, \boldsymbol{A}, \nu^2)} \sim \mathcal{N}(\widehat{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$$

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = (\sigma^{-2}\boldsymbol{\Phi}\boldsymbol{\Phi}^\top + \mathbf{A})^{-1}$$
$$\widehat{\boldsymbol{\theta}} = \sigma^{-2}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{\Phi}\boldsymbol{y} = (\boldsymbol{\Phi}\boldsymbol{\Phi}^\top + \sigma^2\boldsymbol{A})^{-1}\boldsymbol{\Phi}\boldsymbol{y}$$

Getting the predictive distribution

$$p(y|\boldsymbol{x},\mathcal{D}) = \int p(y|\boldsymbol{x},\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D},\sigma^2,\boldsymbol{A},\nu^2)\,d\boldsymbol{\theta} \sim \mathcal{N}(\widehat{y},\sigma_y^2)$$

$$\widehat{y} = \boldsymbol{\phi}^\top(\boldsymbol{x})\widehat{\boldsymbol{\theta}}$$
$$\sigma_y^2 = \sigma^2 + \boldsymbol{\phi}^\top(\boldsymbol{x})\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{\phi}(\boldsymbol{x})$$



Optimizing hyperparameters

σ^2, $\boldsymbol{A}$, ν^2 are optimized by maximizing the **evidence** using EM and simple search

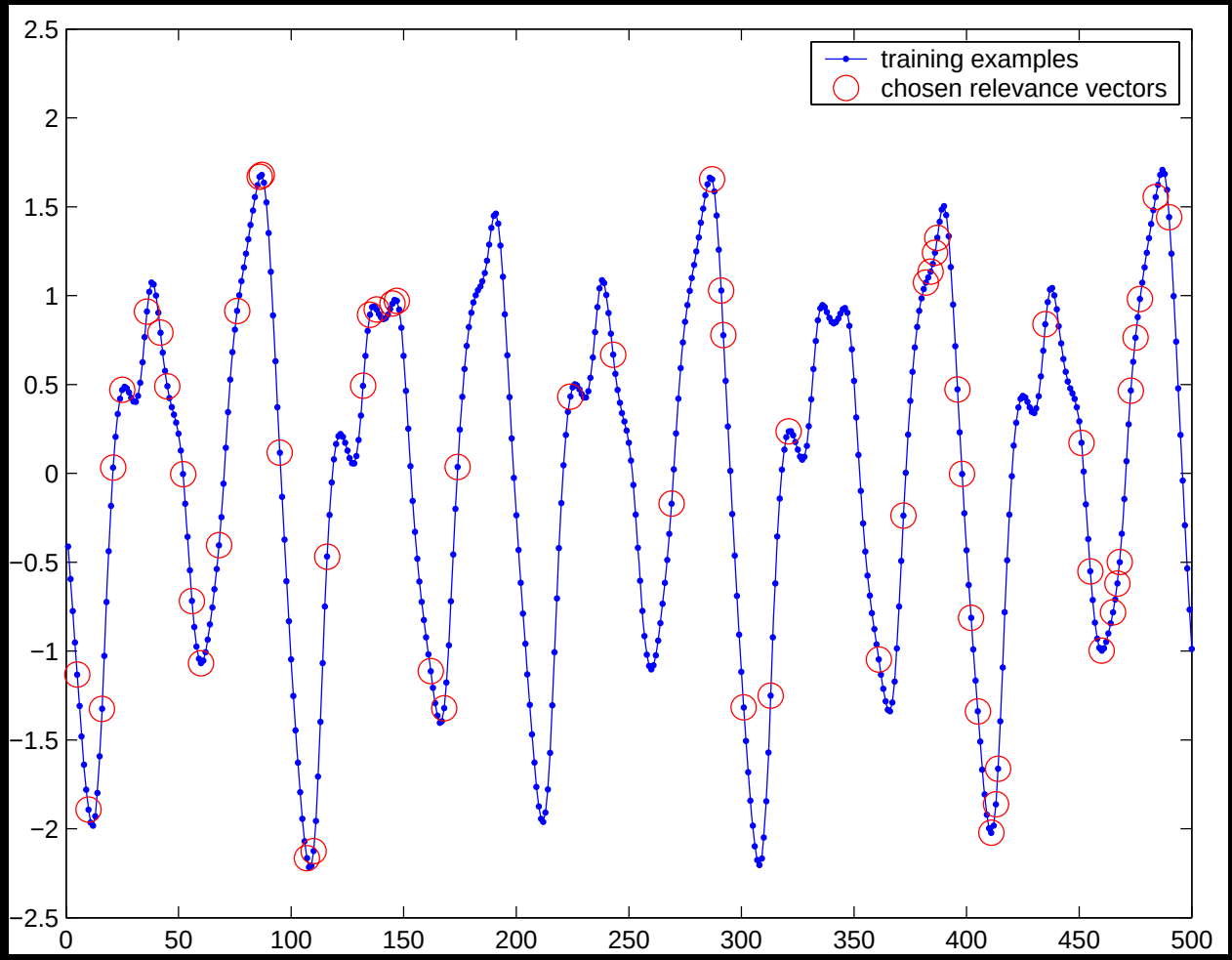
$$p(\mathcal{D}|\sigma^2, \boldsymbol{A}, \nu^2) = p(\mathcal{D}|\boldsymbol{\theta}, \sigma^2, \nu^2)p(\boldsymbol{\theta}|\boldsymbol{A})$$

Literature:

Quiñonero-Candela, J., Girard, A., Larsen, J., Rasmussen, C. E.: "Propagation of Uncertainty in Bayesian Kernel Models - Application to Multiple-Step Ahead Forecasting," ICASSP, vol. 2, pp. 701-704, 2003

Quiñonero-Candela, J., Hansen, L. K.: "Time Series Prediction Based on the Relevance Vector Machine with Adaptive Kernels," ICASSP, pp. 985-988, 2002

Example


training examples
chosen relevance vectors
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
0
50
100
150
200
250
300
350
400
450
500

Outline

- Why Bayesian learning?
- Basic ingredients
- Bayes estimators
- More on selection of priors
- Generalization and bias/variance
- Generalization estimation
- Bayesian model selection
- Discussion of Bayesian framework
- Example of Bayesian learning: RVM
- Bayesian signal detection

Bayesian signal detection model

Literature:

Hansen, L. K., Nielsen, F. ., Larsen, J.; "Exploring fMRI Data for Periodic Signal Components," Artificial Intelligence in Medicine, vol. 25, pp. 25–44, 2002

$$y(n) = \sum_{k=1}^{K} b_k \cdot x_k(n) + \epsilon(n)$$

$$\boldsymbol{y} = \widehat{\boldsymbol{y}} + \boldsymbol{\epsilon} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{\epsilon}$$

- observed signal $\boldsymbol{y} = \{y(n)\}$ is a $N \times 1$ (data $\mathcal{D}$) vector
- $K = 2\kappa$ periodic basis functions $k \in [1; \kappa]$

$$x_{2k}(n) = \cos(k\omega_0 n), \quad x_{2k-1}(n) = \sin(k\omega_0 n)$$

$\boldsymbol{X} = \{x_k(n)\}$ is a $N \times K$ matrix

- b_k linear coefficients $\boldsymbol{b} = \{b_k\}$ is a $K \times 1$ vector
- noise: $\epsilon \sim \mathcal{N}(0, \sigma^2)$

Objective

Estimate unknown fundamental frequency ω_0 and the number of components K

Bayesian learning

- integrate out undesired model parameters $\boldsymbol{\theta} = (\boldsymbol{b}, \sigma^2)$
- select model $\omega_0(m)$, $K(m)$, $m = [1; M]$
- $m = 0$ corresponds to only noise, i.e., $\boldsymbol{X} \equiv 0$.

$$p(\omega_0, K|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\omega_0, K)p(\omega_0, K)}{p(\boldsymbol{y})}$$

Marginal likelihood

$$p(\boldsymbol{y}|\omega_0, K) = \int p(\boldsymbol{y}|\boldsymbol{b}, \sigma^2, \omega_0, K) \cdot p(\boldsymbol{b}, \sigma^2)\, d\boldsymbol{b} d\sigma^2$$
$$= \int (2\pi\sigma^2)^{N/2} \exp\left(-\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2/2\sigma^2\right) \cdot p(\boldsymbol{b}, \sigma^2)\, d\boldsymbol{b} d\sigma^2$$

Conjugate prior: normal-inverse-gamma

$$p(\boldsymbol{b}, \sigma^2|a, d, K, \boldsymbol{m}, \boldsymbol{V}) = \frac{(a/2)^{d/2} \cdot (\sigma^2)^{-(d+K+2)/2}}{(2\pi)^{K/2} \cdot \det \boldsymbol{V}^{1/2} \cdot \Gamma(d/2)} \cdot$$

$$\exp\left(-(\boldsymbol{b} - \boldsymbol{m})^\top (2\sigma^2 \boldsymbol{V})^{-1} (\boldsymbol{b} - \boldsymbol{m}) - a/2\sigma^2\right)$$

Priors

$$p(\boldsymbol{b}|a,d,K,\boldsymbol{m},\boldsymbol{V}) = \int p(\boldsymbol{b},\sigma^2|a,d,K,\boldsymbol{m},\boldsymbol{V})\,d\sigma^2$$
$$\sim \mathcal{T}(\boldsymbol{m}, a\boldsymbol{V}/(d-2))$$

$$p(\sigma^2|a,d) = \int p(\boldsymbol{b},\sigma^2|a,d,K,\boldsymbol{m},\boldsymbol{V})\,d\boldsymbol{b} \sim \mathcal{IG}(a/(d-2))$$

- mean of noise variance $a/(d-2) = \widehat{\sigma}_y^2 = \boldsymbol{y}^\top\boldsymbol{y}/N$
- $d = 3$ is smallest value for which prior is finite, hence "weak"
- $\boldsymbol{m} = \boldsymbol{0}$ for no prior assumption of mean amplitude of periodic components

Priors

- $\boldsymbol{V} = v\boldsymbol{I}$ for simplicity

$$E_{prior}[\boldsymbol{y}^\top \boldsymbol{y}]/N = Tr[\boldsymbol{X}\boldsymbol{X}^\top E_{prior}[\boldsymbol{b}\boldsymbol{b}^\top]]/N$$
$$= v \cdot a/(d-2) \cdot Tr[\boldsymbol{X}\boldsymbol{X}^\top]/N$$

If $E_{prior}[\boldsymbol{y}^\top \boldsymbol{y}]/N = \widehat{\sigma}_y^2 = a/(d-2)$ then $v = N \cdot Tr[\boldsymbol{X}\boldsymbol{X}^\top]^{-1}$

Marginal likelihood

posterior is NIG thus the marginal likelihood is its normalization integral

$$p(\boldsymbol{y}|\omega_0, K) = \left(\frac{\det \boldsymbol{V}_P \cdot a^d}{\det \boldsymbol{V} \cdot a_P^{d_P} \cdot \pi^N} \right)^{1/2} \frac{\Gamma(d_P/2)}{\Gamma(d/2)}$$

- $\boldsymbol{V}_P^{-1} = \boldsymbol{V}^{-1} + \boldsymbol{X}^\top \boldsymbol{X}$
- $\boldsymbol{m}_P = \boldsymbol{V}_P(\boldsymbol{V}^{-1}\boldsymbol{m} + \boldsymbol{X}^\top \boldsymbol{y})$
- $a_P = a + \boldsymbol{m}^\top \boldsymbol{V}^{-1}\boldsymbol{m} + \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{m}_P^\top \boldsymbol{V}_P^{-1}\boldsymbol{m}_P$
- $d_P = d + N$

Model selection

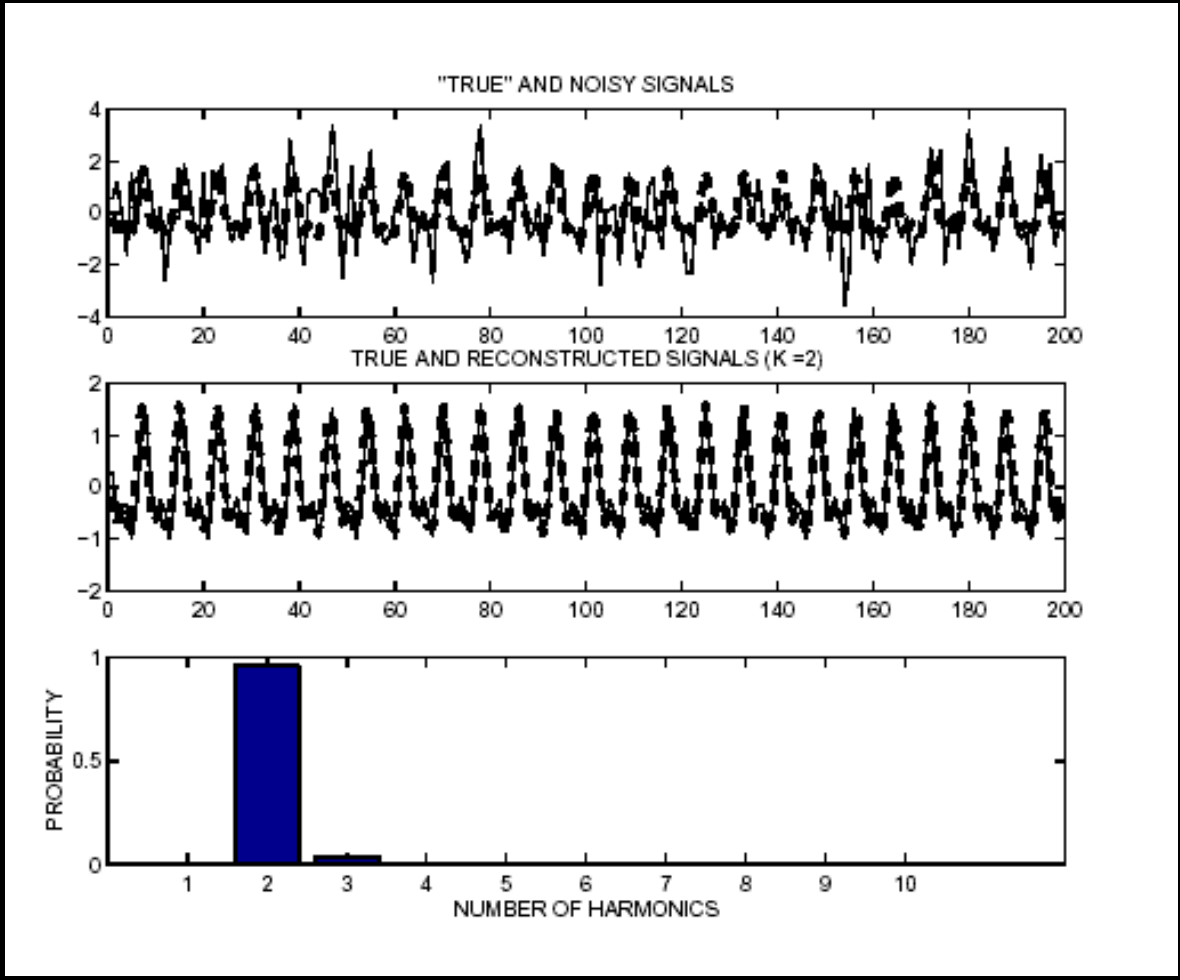
uniform models prior $p(\omega_0(m), K(m)) = 1/(M+1)$

$$p(m|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|m)}{\sum_{m=0}^{M} p(\boldsymbol{y}|m)}$$

with

- $p(\boldsymbol{y}|m) = p(\boldsymbol{y}|\omega(m), K(m)),\ m > 0$
- $p(\boldsymbol{y}|0) = p(\boldsymbol{y}|\omega, K)$ with $\boldsymbol{X} = 0$

Example


"TRUE" AND NOISY SIGNALS
TRUE AND RECONSTRUCTED SIGNALS (K =2)
PROBABILITY
NUMBER OF HARMONICS

Suggestions for a roadmap

- More research on the evaluation of the learning process. Account for all variation – also data set variability
- Development of better and easy to communicate approximation schemes
- More research on online learning in a non-stationary switching dynamics settings
- Bayes does not tell you anything about the domain exterior to the model - hence, more focus on integrating the data representation, feature selection, and preprocessing steps
- Systems interact with other systems and humans – model the man in the loop, model irrationality

Wrap up

- Bayesian learning combines all available knowledge in principled way
- Ingredients: variable, data, model, prior, loss
- Bayes is optimal in admissibility/generalization sense
- Bayes framework is complete as it offers model selection and confidence
- Robustness needs to be tested, model mis-specifications can cause arbitrary errors