Ensemble Learning

based on the similarly named article by Harri Lappalainen and James W. Miskin

- Bayes' Theorem
- Overfitting
- Ensemble Learning
- Priors and Hyperpriors

Tapani Raiko 16.11.2000

Bayes' Theorem

- One can evaluate the conditional probability distributions for the unknown parameters, θ, given the set of observed quantities, x.

$$p(x, \theta) = p(x|\theta)p(\theta) = p(\theta|x)p(x)$$

- $p(x)$ is constant, when comparing different θ.

$$\Rightarrow p(\theta|x) \propto p(x|\theta)p(\theta)$$

- The prior distribution $p(\theta)$ contains knowledge of parameters before observations.
- The posterior probability density $p(\theta|x)$ can be used for many purposes.

Simple Bayesian example

- Model:

$$y = \theta x + n \Rightarrow n = y - \theta x$$

- Prior distributions:

$$p(\theta) = N(0,1), \;\; p(n) = N(0,\epsilon)$$

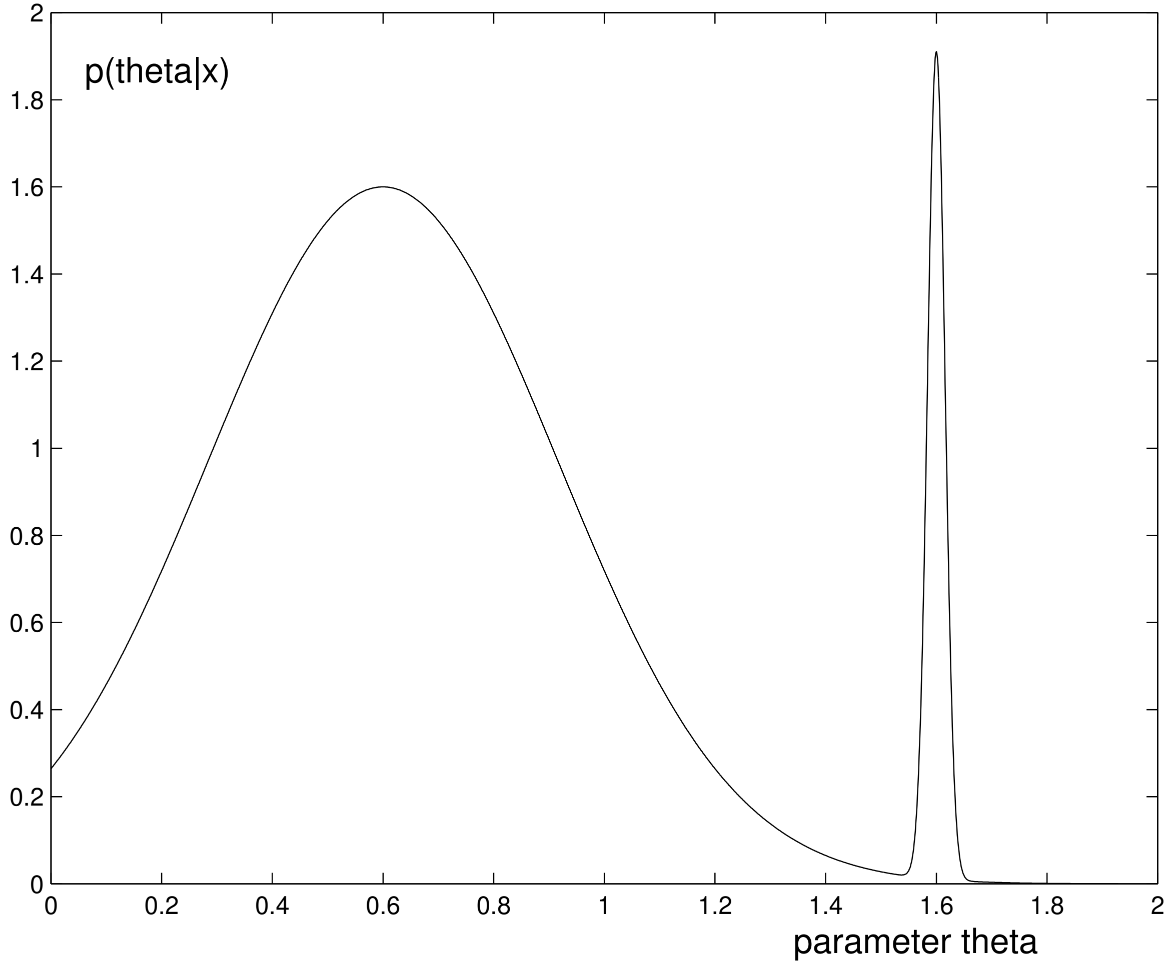
- Posterior distribution:

$$p(\theta|x) \propto p(x|\theta)p(\theta) \propto e^{-\frac{(y-\theta x)^2}{2\epsilon}} e^{-\frac{\theta^2}{2}}$$

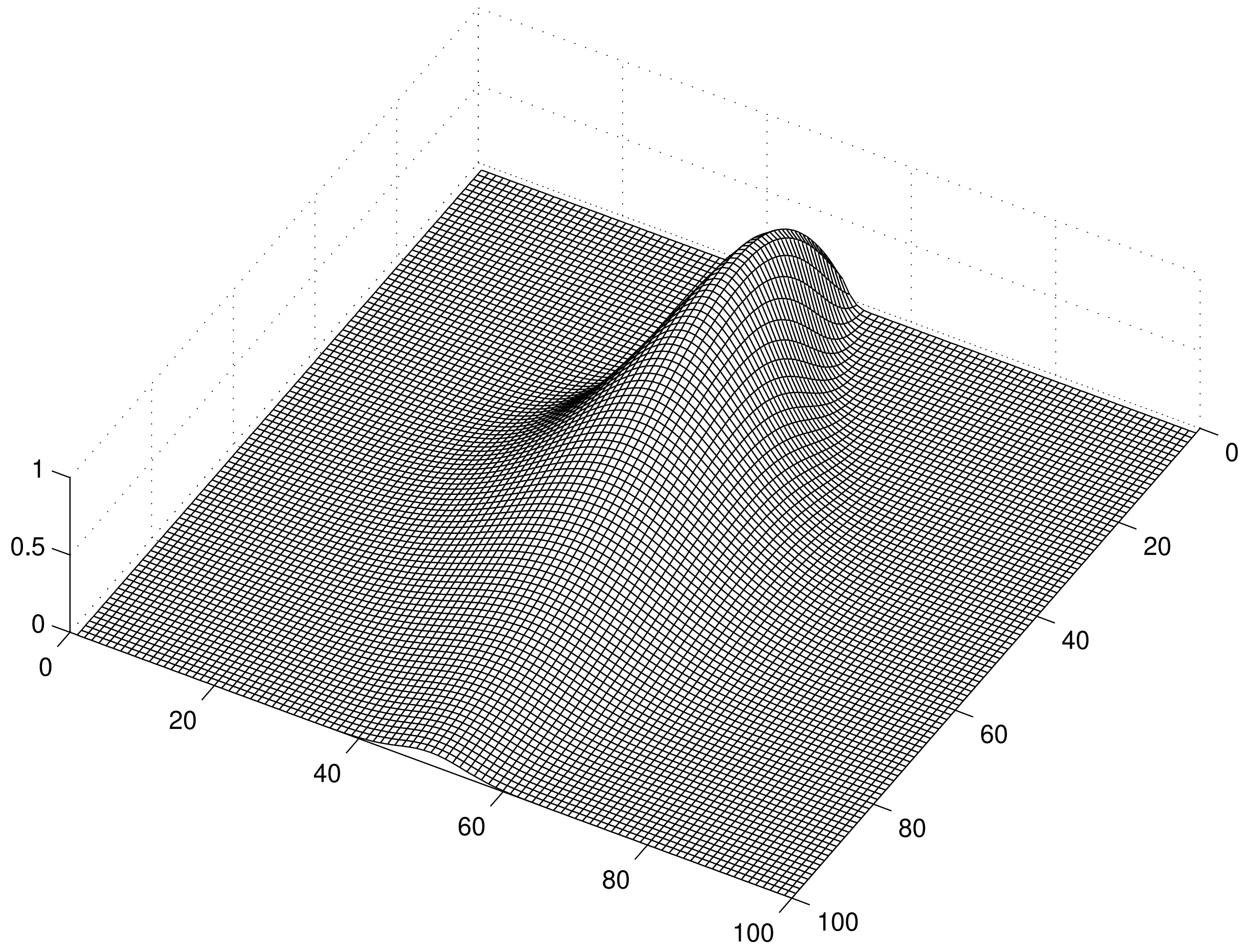
- We don't select a single point θ, instead, we consider all of them.

- "Best" θ can be selected (if needed) by marginalizing: $\bar{\theta} = \int \theta p(\theta|x) d\theta$

What Is Overfitting?



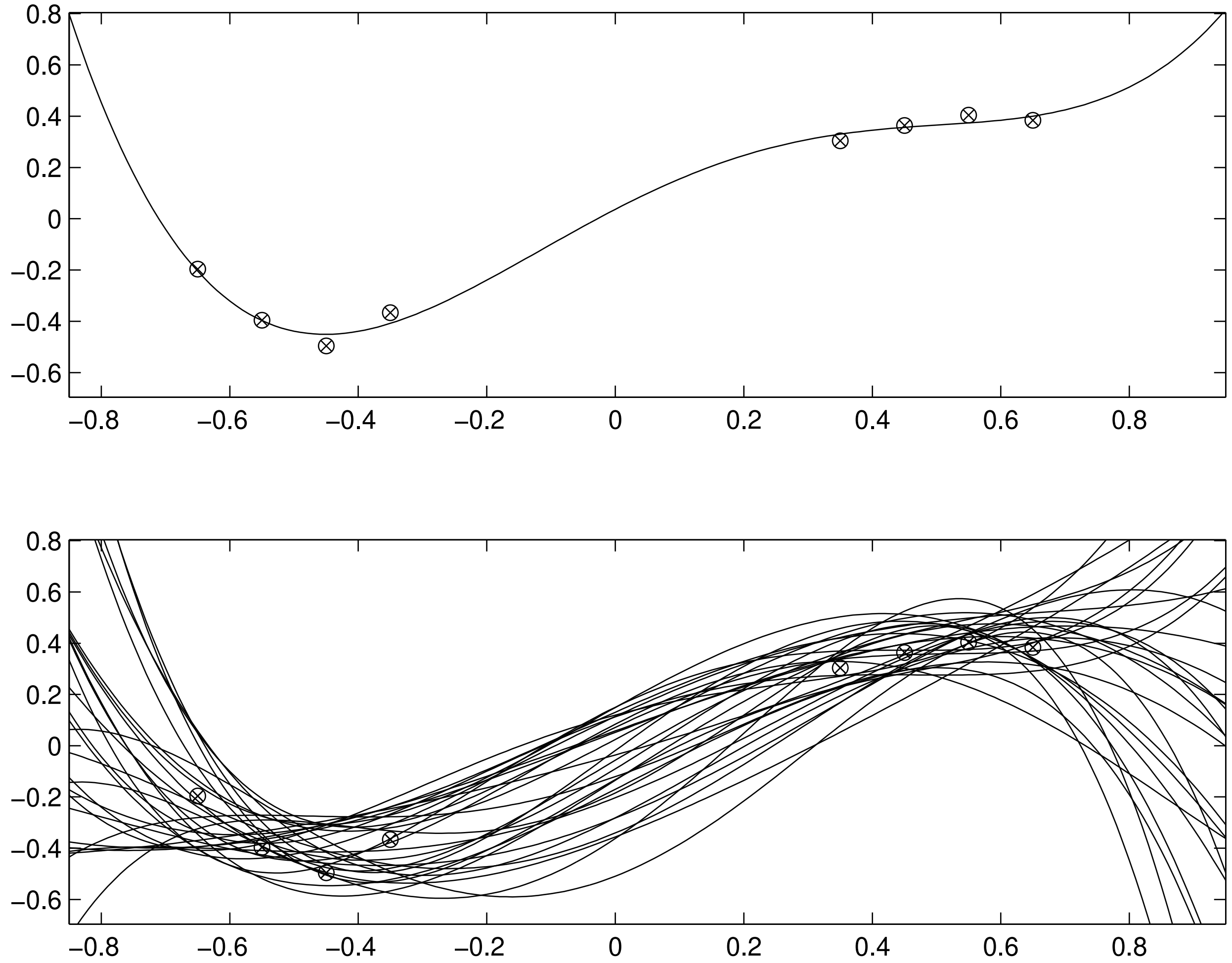
Multidimensional Case



Approximations of Posterior pdf

- Marginalization requires integrating over posterior density, which is too difficult without approximations.
- Point estimate would lead to overfitting.
- Stochastic approximations
  - Draw samples from posterior pdf and average over samples.
- Parametric approximations
  - Use simpler distribution to approximate the posterior pdf.

Fitting vs. Sampling of Fifth-Order Polynomials



Kullback-Leibler information

- is a distance measure between two distributions p and q.

$$\begin{aligned} D(q(\theta|x)||p(\theta|x)) &= \int q(\theta|x) \ln \frac{q(\theta|x)}{p(\theta|x)} d\theta \\ &= \int q(\theta|x) \ln \frac{q(\theta|x)}{p(x,\theta)} d\theta + \ln p(x) \end{aligned}$$

- Note that KL-information is asymmetric.
- KL-information is sensitive to "mass" rather than single values of the distribution.

Ensemble Learning

- is a method for parametric approximation of posterior pdf.
- Kullback-Leibler information is used to measure the misfit between the actual posterior distribution p and its approximation q.
- Term $\ln p(x)$ is constant and can be left out.
- After a parametric form for the ensemble is fixed, the training is tractable. (e.g. assuming a product of Gaussians)
- An alternative method is to assume separable form for the ensemble. The distributions may then be found by functional minimisation, which is not always tractable.

Model Selection

- Same principle applies even when considering an ensemble of multiple models.

$$p(H|x) = \frac{p(x|H)p(H)}{p(x)} \approx \frac{e^{-C_\theta(x|H)}P(H)}{\sum_{H'} e^{-C_\theta(x|H')}P(H')},$$

  where $C_\theta(x)$ is the KL-information minus the constant.

$$\begin{aligned} C_\theta(x) = D(q(\theta|x)||p(\theta|x)) - \ln p(x) &= \int q(\theta|x) \ln \frac{q(\theta|x)}{p(x,\theta)} d\theta \\ &= \int (q \ln q - q \ln(x|\theta) - q \ln p(\theta)) d\theta \end{aligned}$$

Example

$$x = f(\theta) + n$$

$$p(\theta) = N(y_e, y_{var})$$

$$p(n) = N(0, \sigma^2)$$

$$q(\theta|x) = \frac{e^{-\frac{(\theta-\bar{\theta})^2}{2\tilde{\theta}}}}{\sqrt{2\pi\tilde{\theta}}}$$

$$p(x|\theta) = \frac{e^{-\frac{(x-f(\theta))^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$$

$$\begin{aligned} C_\theta(x) &= \int (q \ln q - q \ln(x|\theta) - q \ln p(\theta)) d\theta \\ &= -\frac{1}{2} \ln 2\pi e\tilde{\theta} + \frac{(\bar{\theta} - \theta_e)^2 + \tilde{\theta}}{2\theta_{var}} \\ &\quad + \frac{(x - \bar{f}(\bar{\theta}, \tilde{\theta}))^2 + \tilde{f}(\bar{\theta}, \tilde{\theta})}{2\sigma^2} \end{aligned}$$

Connection to Coding

- There is a close connection between coding and probabilistic framework.
  Optimal code length of x is $L(x) = -\log P(x)$ bits.
- In order to be able to encode data compactly one has to find a good model for it.
  Message will have two parts: model θ and data x.
- It can be shown, that:

$$L(x) = C_\theta(x) - \log dx,$$

where dx is the required accuracy of data.

Construction of Probabilistic Models

- We shall see how a generative model turns into a probabilistic model.

$$x(t) = f(y(t), \theta) + n(t)$$

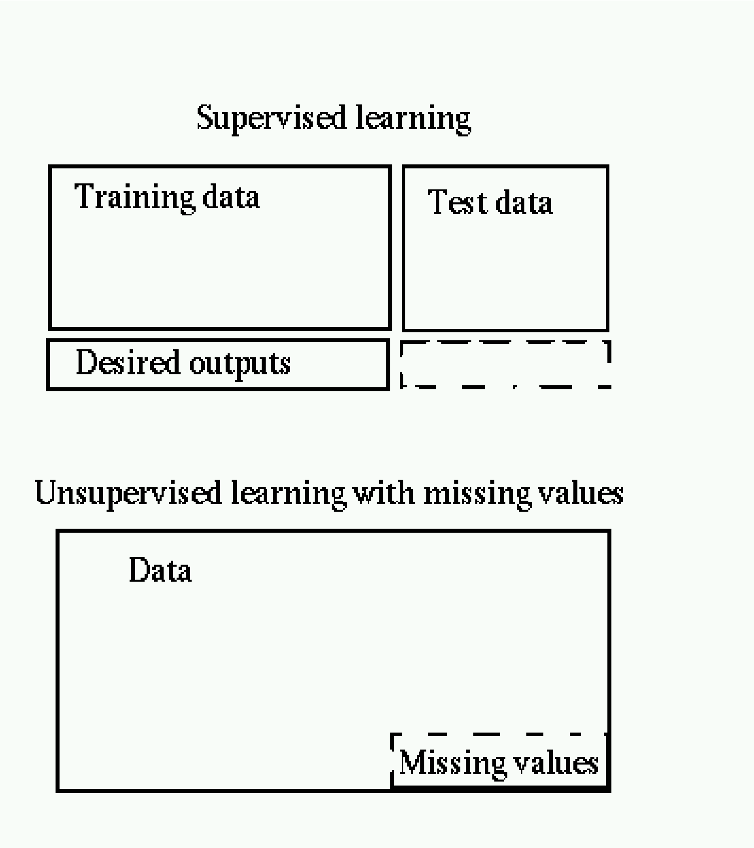
- If $y(t)$ and θ are given, then $x(t)$ has the same distribution as $n(t)$ except that it is offset by $f(y(t), \theta)$.

$$n(t) \sim N(0, \sigma^2) \Rightarrow x(t) \sim N(f(y(t), \theta), \sigma^2)$$

- This gives us the joint density of all the variables.

$$p(x, y, \theta, \sigma) = p(y, \theta, \sigma) \prod_t p(x(t)|y(t), \theta, \sigma)$$

Missing Values



- Treatment of missing values is possible as long as the model defines the joint density of all the variables in the model.

Priors

- Parameters and model structures need to have a prior probability.
- In coding terms, the prior means the aspects of the encoding which the sender and the receiver have agreed upon prior to the transmission.
- The term “uninformative prior” means that the prior gives as little information as possible. Unfortunatedle they tend not to be normalisable and cannot be used.
- Use of priors is the essence of the Bayesian method.

Hyperparameters

- It is reasonable to assume that for example all biases in an MLP network have a similar distribution.
- Hyperparameter controls the distribution of a set of parameters.
- This represents our structural knowledge of the model.
- Example:

$$
\begin{aligned}
p(\theta_i) &= N(\theta_e, e^{2\theta_{var}}) \\
p(\theta_e) &= N(0, 100) \\
p(\theta_{var}) &= N(0, 100)
\end{aligned}
$$

Why?

- World makes sense to us because we don't observe only photons and changes in air pressure etc. Instead, we observe objects and people and hear them speak and sing. We can put lots of data into a computer and it would be nice, if it could make some sense of it.
- Some near future applications could be speech and image recognition, video compression, artificial intelligence (in some compact domain) and anything which is based on MLP-networks or ICA.
- Non-Linear Factoral Analysis.

Summary

- Ensemble Learning makes Bayesian method tractable.
- The problem of overfitting is removed.
- Model comparison is straightforward.
- Combines supervised and unsupervised learning.
- The approximation of the posterior distribution assumes some degree of factorisation. Also some fixed functional form of the factors is usually assumed.
- Computational complexity is greater, but not overwhelming, compared to traditional methods.