Probabilistic Models in Machine Learning

Gunnar Rätsch\textsuperscript{1}, Cheng Soon Ong\textsuperscript{1,2}, Petra Philips\textsuperscript{1}

\textsuperscript{1} Friedrich Miescher Laboratory, Tübingen
\textsuperscript{2} MPI for Biological Cybernetics, Tübingen

Vorlesung WS 2007/2008
Eberhard-Karls-Universität Tübingen

20 November 2007

http://www.fml.mpg.de/raetsch/lectures/amssa07
Lecture Outline

Basics
- The Probabilistic Model
- Some Probability and Statistics
- Generative vs. Discriminative Approach

Maximum Likelihood
- Idea
- Gaussian Distributions
- ML for Gaussians
The ‘Learning World’:

**Data:**  \( \{(x,y)\}_{d=1}^{D}, x\in \{\text{natural, plastic}\}, y\in \{-1, 1\} \)

**Objective:** Given new \( x \), predict \( y \)

**Problem:** The world is incomplete and uncertain!
The ‘Learning World’:  

Data: \( \{(x,y)\} = 1 \left\{ x_{\in}, y_{\in} \{ \pm 1 \} \right\} \)

Objective: Given new \( x \), predict \( y \)

Probabilistic Assumption: All data is generated by the same hidden probabilistic source!
The ‘Learning World’:

Data: \( \{(x,y)\}_{1}^{d} x_{d}, y \in \{-1,1\} \)

Objective: Given new \( x \), predict \( y \)

**Probabilistic Assumption:** All data is generated by the same hidden probabilistic source!
The ‘Learning World’:

Data: \[ \{(x, y)\}_{1}^{d}, y \in \{\pm 1\} \]

Objective: Find ’best’ \[ y = h(x) \]

Assumption: Sample \[ S = (x_1, y_1), \ldots, (x, y) \sim p(x, y) \]

**Problem:** \( p(x, y) \) is unknown!
Given:

Hypothesis class: \( \{ h \mid h : d \to y \} \) (e.g. \( y = \{ \pm 1 \} \))

Loss: \( l(y, h(x)) \) (e.g. 0-1 loss \([y \neq h(x)]\))

Sample: \( S = (x_1, y_1), \ldots, (x,y) \sim p(x, y) \)

Objective: Minimize the true (expected) loss – “generalization error”

\[
h^* = \min_{h \in H} L(h) \quad \text{with} \quad L(h) := \mathbb{E}_x l(., h())
\]

Problem: Only have a data sample available, \( p(x, y) \) is unknown!

This is a very general setup! By varying \( h \) and \( l \) one arrives at classification, regression, clustering.

One example is classification.
Given:
Hypothesis class: $= \{ h \mid h : d \rightarrow Y \}$ (e.g. $y = \{\pm 1\}$)

Loss: $l(y, h(x))$ (e.g. 0-1 loss $[y \neq h(x)]$)

Sample: $S = (x_1, y_1), \ldots, (x, y) \sim p(x, y)$

Objective: Minimize the true (expected) loss – “generalization error”

$$h^* = \min_{h \in \text{red}} L(h) \text{ with } L(h) := \mathbb{E}_x l(, h())$$

Problem: Only have a data sample available, $p(x, y)$ is unknown!

This is a very general setup! By varying and $l$ one arrives at classification, regression, clustering.

One example is classification.
**Solution**

**Given:**
- **Hypothesis class:** \( \{ h \mid h : d \to \{ \pm 1 \} \} \)
- **Loss:** \( l(y, h(x)) \) (e.g. 0-1 loss \([y \neq h(x)]\))
- **Sample:** \( S = (x_1, y_1), \ldots, (x, y) \sim p(x, y) \)

**Objective:** Minimize the true (expected) loss — “generalization error”

\[
    h^* = \min_{h \in \text{red}} L(h) \quad \text{with} \quad L(h) := \mathbb{E}_{(x, y)} l(y, h(x))
\]

**Problem:** Only have a **given** data sample available, \( p(x, y) \) is unknown!

**Solution:** Estimate \( p(x, y) \) from \( S \). Then minimize

\[
    \hat{h} = \hat{L}(h) := \hat{\mathbb{E}}_{(x, y)} l(y, h(x)).
\]
Can we estimate any \( p(x, y) \) well algorithmically?

- No free lunch!

How does 'the approximation' of \( p(x, y) \) influence the generalization error of \( \hat{h} \)?

- Occam’s razor: good approximation on sample + simple model \( \Rightarrow \) good generalization.

- Empirical distribution (ERM): statistical learning theory quantifies how generalization is influenced by choice of \( \ell \).

Can we do estimation and optimization efficiently?

- Sometimes, depending on our model.
Least Squares

Linear regression

\[ y = \mathbf{w}^\top \mathbf{x} + \varepsilon \]

Best Unbiased Linear Estimator

\[
\min_{\mathbf{w}} \sum_{i=1}^{n} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2
\]

Gradient

\[
\nabla_{\mathbf{w}} = \sum_{i=1}^{n} (y_i - \mathbf{w}^\top \mathbf{x}_i) \mathbf{x}_i^\top
\]

Solution

\[
\mathbf{w}_{LS} = (X^\top X)^{-1} X^\top y
\]
Regularized Least Squares

Linear regression

\[ y = w^\top x + \varepsilon \]

Bias the weights to be “small”

\[
\min_w \sum_{i=1}^{n} (y_i - w^\top x_i)^2 + \lambda w^\top w
\]

Gradient

\[
\nabla_w = \sum_{i=1}^{n} (y_i - w^\top x_i) x_i^\top + \lambda w
\]

Solution

\[
w_{RLS} = (X^\top X + \lambda I)^{-1} X^\top y
\]
Reading

- Probability books: ????.
- Machine Learning: ???.
- Theory of Classification: ??

Online Resources

**Probability**

**Random Variable** is a function that maps outcomes of random experiments to numbers.

**Probability** is the fraction of times an event occurs.

**Lazy notation**

\[ p(x) \] denotes the probability that random variable \( X \) takes value \( x \), i.e. \( p(X = x) \).

For continuous variables, \( p(x) \) is technically called the probability density.

**Note that these are not definitions!**
Joint Probability
The entry of both values jointly.

\[ p(X = x_i, Y = y_i) = \frac{n_{ij}}{N} \]

Marginal Probability
The sum over a row or column.

\[ p(X = x_i) = \frac{c_i}{N} \]

Conditional Probability
The fraction of a row or column in a particular cell.

\[ p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i} \]
Expectation  For discrete distributions, defined as the average value of $f(x)$ weighted by the probabilities

$$E_X f(X) = \sum_i f(x_i)p(X = x_i).$$

Conditional Expectation  Expectation w.r.t. the conditional distribution

$$E_X [f(X)|Y] = \sum_i f(x_i)p(X = x_i|Y).$$
Useful relations

**Factorization** If \( X \) and \( Y \) are independent events then

\[
p(X = x_i, Y = y_j) = p(X = x_i)p(Y = y_j)
\]

**Marginalization (Sum rule)**

\[
p(X = x_i) = \sum_j p(X = x_i, Y = y_j).
\]

**Product Rule**

\[
p(x, y) = p(y|x)p(x) = p(x|y)p(y).
\]

**Bayes’ Rule**

\[
p(y|x) = \frac{p(x|y)p(y)}{p(x)}
\]

posterior \( \propto \) class cond. \( \times \) prior
**Classification:** The function which minimizes the expected 0-1 loss is the one which assigns to each \( x \) the class with maximal posterior

\[
h(x) := \begin{cases} 
1, & \text{if } p(Y = 1|x) \geq p(Y = -1|x) \\
-1, & \text{otherwise}
\end{cases}
\]

(This easily extends to the multiclass case and to arbitrary loss-matrix.)

**Regression:** The function which minimizes the mean square error is conditional expectation w.r.t the posterior

\[
h(x) := \mathbb{E}(y|x).
\]

What we want is thus \( p(y|x) \).
Generative vs Discriminative

**Generative approach**
Models $p(x, y) = p(x|y)p(y)$. Uses Bayes’ rule to infer

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}.$$

**Discriminative approach**
Models $p(y|x)$ directly and takes max.
Examples

Generative

- Mixtures of Gaussians
- Hidden Markov Models
- Bayesian Networks
- Graphical Models
- …

Discriminative

- SVM
- Least Squares Regression
- Regularized Least Squares Regression
- …
Maximum Likelihood Approach

- Principled approach for density estimation when given data samples.
- Assume a parameterized class of densities:
  \[ p(x) \in \{ p(x|\theta) | \theta \in \Theta \} \].
- The likelihood function: models probability of observed data for each parameter:
  \[ l_x(\theta) = p(x|\theta) \].
- Idea: **After** seeing a data sample \( D = x_1, x_2, \ldots, x_n \) (usually iid) pick parameter which is most likely to have generated this sample (parameter estimation):
  \[ \theta^{*}_{ML} = \text{argmin}_{\theta} l_D(\theta) \].
IID Data: $D = x_1, x_2, \ldots x_n$

$$l_D(\theta) = \prod_{i} p(x_i | \theta).$$

Negative Log Likelihood

Usually minimize $-\log p(D | \theta)$.

- Logarithm is monotonic
- Simplifies maths: IID assumption makes log a sum (derivatives termwise)
- Better numerical behaviour
Gaussian Distribution (1)

1 dimensional

\[
(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp \left( -\frac{1}{2\sigma^2}(x - \mu)^2 \right)
\]

D dimensional

\[
(x|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left( -\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right)
\]
Gaussian Distribution (2)

D dimensional

\[
(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp \left( -\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu) \right)
\]

Negative Log Likelihood

\[
-\log p(x) = \frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu) - \frac{1}{2} \log |\Sigma| - \frac{D}{2} \log 2\pi
\]

Important Properties

- All marginals and conditionals are again Gaussian.
- Sufficient statistics are mean and variance.
- Part of the exponential family distributions.
Likelihood Function

\[ p(y|x) = p(y|x, \theta); \quad l_x(\theta) = p(y|x, \theta) \sim (\mu(\theta), \sigma(\theta)^2) \]

Linear Regression with Gaussian Noise

\[ y = w^\top x + \varepsilon, \]

where \( \varepsilon \sim (0, \sigma^2) \).

Gaussian noise

Using the fact that \( \varepsilon \) is Gaussian, then \( y_i \sim (w^\top x_i, \sigma^2) \), and hence

\[ p(y|x, w) = \prod_{i=1}^{n} \left( y_i - w^\top x_i, \sigma^2 \right) \]
Optimization

Maximum Likelihood with Gaussian Noise

\[ p(y|x, w) = \prod_{i=1}^{n} (y_i - w^\top x_i, \sigma^2) \]

Negative Log Likelihood

\[ \min_w - \log p(y|x, w) = \sum_{i=1}^{n} (y_i - w^\top x_i)^2 \]

Least Squares

Observe that the negative log likelihood is the least squares problem.

Solution

\[ w_{ML} = (X^\top X)^{-1} X^\top y \]
Maximum Likelihood

We have minimized the negative log likelihood,
\[- \log p(y|x, w)\]
to obtain the weight vector \( w_{ML} \).

Recall

We have a probabilistic model, and hence we can make a predictive distribution for a new value \( x_t \).

In other words

For each test example \( x_t \), we get a distribution
\[ y_t \sim (w_{ML}^\top x_t, \sigma^2). \]
Regression: Variance
Prior

Bayes rule

\[ p(w | x, y) = \frac{p(y | x, w)p(w)}{p(y | x)} \]

posterior $\propto$ likelihood $\times$ parameter prior

Using background information

We can include prior knowledge about how we would like the weight vector, for example that it is multivariate Gaussian:

\[ w \sim (0, \Sigma^{-1}). \]

i.e.

\[ p(w) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp \left( -\frac{1}{2} w^\top \Sigma^{-1} w \right). \]
Maximum a Posteriori

Idea

Instead of finding $w$ that maximizes likelihood $p(y|x,w)$, we find the $w$ that maximizes the posterior $p(w|x,y)$.

Posterior

$$p(w|x,y) \propto \exp \left( -\frac{1}{2\sigma^2} (y - w^\top X)^\top (y - w^\top X) \right) \exp \left( -\frac{1}{2} w^\top \Sigma^{-1} w \right)$$

Negative Log Posterior

$$-\log p(w|x,y) = -\frac{1}{2\sigma^2} (y-w^\top X)^\top (y-w^\top X) - \frac{1}{2} w^\top \Sigma^{-1} w$$

Regularized Least Squares

The negative log posterior is the regularized least squares problem.
How MAP Works

likelihood prior/posterior data space
Maximum Likelihood with Gaussian posteriors for linear regression with Gaussian noise is the least squares solution.

Maximum a posteriori with a Gaussian prior for the same problem is the regularized least squares solution.
The posterior probability for the positive class is given by

\[ p(Y = 1|x) = \frac{1}{1 + \exp(-a)} = \text{sig}(a) \]

where

\[ a = \ln \frac{p(x|Y = 1)p(Y = 1)}{p(x|Y = -1)p(Y = -1)}. \]

**Sigmoid Function**

\[ \text{sig}(a) = \frac{1}{1 + \exp(-a)} \]

(“S” shaped).
ML for Class Conditionals

Maximum Likelihood: Assume Gaussian class conditional densities with equal covariances and solve maximum likelihood with constraint

\[ p(x|Y = 1) \sim (\mu_+, \Sigma^{-1}); \quad p(x|Y = -1) \sim (\mu_-, \Sigma^{-1}). \]

Linear Term in \( x \): From the results in the previous slide, the posterior is

\[ p(Y = 1|x) = \text{sig}(w^\top x + w_0), \]

where

\[ w = \Sigma^{-1}(\mu_+ - \mu_-) \]

\[ w_0 = -\frac{1}{2} \mu_+^\top \Sigma^{-1} \mu_+ + \frac{1}{2} \mu_-^\top \Sigma^{-1} \mu_- + \ln \frac{p(Y = 1)}{p(Y = -1)} \]

Linear Classifier: \( w^* = \arg\max_w \text{sig}(y_i(w^\top x + w_0)) \)

\[ = \arg\min_w (1 + \exp(-y_i(w^\top x + w_0))). \]
Logistic Regression

- Logistic regression solves a **classification** task.
- It assumes a sigmoid function acting on a linear model, i.e. it assumes a logistic loss
  \[
  \ell(f(x_i), y_i) := \log(1 + \exp(-y_i f(x_i)))
  \]
- Can solve for the maximum likelihood (unregularized) or maximum a posteriori (regularized) weights.
Probabilistic view provides different tools for analysis

Bayes rule gives the relationship between the class conditionals and the posterior.

From the probabilistic model, one can do both generative and discriminative methods via maximum likelihood.

Maximum likelihood approach for posteriors or class conditionals recovers least squares regression and logistic regression.

MAP can be seen as regularized ML estimation.