Advanced Methods for Sequence Analysis

G. Rätsch[1], C.S. Ong[1,2] and P. Philips[1]

[1] Friedrich Miescher Laboratory, Tübingen
[2] Max Planck Institute for Biological Cybernetics, Tübingen

Vorlesung WS 2006/2007
Eberhard Karls Universität Tübingen

17 January 2007

http://www.fml.mpg.de/raetsch/lectures/amsa

What is a Model?



Lecture Outline

- Linear Least Squares Regression
  - Note that can apply kernel trick
  - Note that problem is convex and differentiable
- Probability Whirlwind Tour
  - Notation
  - Likelihood, Prior, Posterior
- Gaussian Distribution
  - Maximum Likelihood
  - Maximum a Posteriori
- Logistic Regression

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

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

Bias the weights to be "small"

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

Gradient

$$\nabla_{\mathbf{w}} = \sum_{i=1}^{n} (y_i - \mathbf{w}^\top \mathbf{x}_i)\mathbf{x}_i^\top + \lambda \mathbf{w}$$

Solution

$$\mathbf{w}_{RLS} = (X^\top X + \lambda \mathbf{I})^{-1} X^\top y$$

Probabilistic View (1)

Model the machine learning task as the joint probability

$$p(\mathbf{x}, y).$$

example $\mathbf{x}_i \in \mathcal{X}$ for example a set of amino acid sequences.

label $y_i \in \mathcal{Y}$ for example whether the sequence is an alpha helix, beta sheet of coil.

Training Data Data consisting of examples and associated labels, which are used for training the machine.

Testing Data Data consisting only of examples, used for generating predictions.

Predictions Output of the trained machine.

Probabilistic View (2)

Consider the problem of linear regression,

$$y = \mathbf{w}^\top \mathbf{x} + \varepsilon.$$

1. How are examples represented?
2. How are labels represented?
3. What does the machine estimate?
4. What does the machine output?

Reading

- Probability books: Mackay [2003], Jacod and Protter [2004], Dudley [2002], Shiryayev [1984].
- Machine Learning: Bishop [2006], Duda et al. [2001], Rasmussen and Williams [2006].
- Online Resources
  - http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/pdf.html
  - http://www.inference.phy.cam.ac.uk/mackay/itila/book.html

Probability (1)

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, Marginal, Conditional (1)

Joint Probability

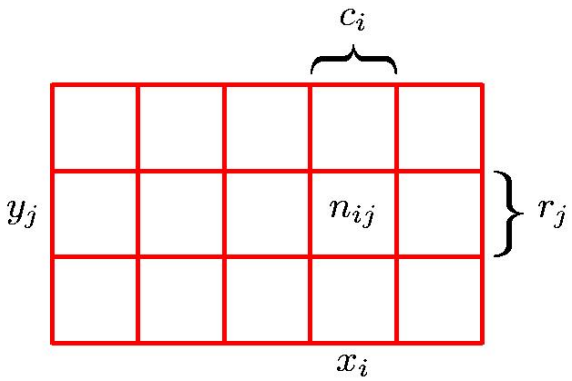
The entry of both values jointly.

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

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.

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

Joint, Marginal, Conditional (2)

Joint Probability

For two random variables X and Y, the probability that $X = x$ and $Y = y$ is (lazily) written as $p(x, y)$ and is called the joint probability.

Marginal Probability

The probability that X takes the value x irrespective of the value of random variable Y is (lazily) written as $p(x)$ and is called the marginal probability.

Conditional Probability

If we consider only the instances where $X = x$, then the fraction of instances for which $Y = y$ is written (lazily) as $p(y|x)$ and is called the conditional probability.

Useful relations

Factorization If X and Y are independent events then

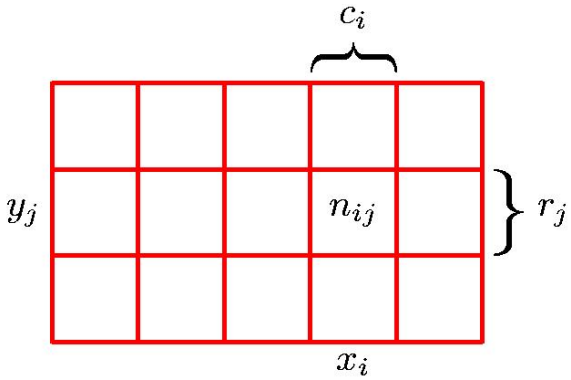
$$p(X = x_i, Y = y_j) = p(X = x_i)p(Y = y_j)$$

Marginalization (Sum rule)

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

Product Rule

$$p(x, y) = p(y|x)p(x).$$



Bayes' Rule

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

posterior $\propto$ likelihood $\times$ prior

Probabilistic View (3)

Consider the problem of linear regression,

$$y = \mathbf{w}^\top \mathbf{x} + \varepsilon.$$

1. How are examples represented?
   - $\mathbf{x}$ or in feature space $\Phi(\mathbf{x})$.
2. How are labels represented?
   - y
3. What does training estimate?
   - $p(\mathbf{w}|\mathbf{x}, y)$
4. What does the machine output?
   - $p(y|\mathbf{w}, \mathbf{x})$

Generative vs Discriminative (1)

Generative approach

- Use model to generate examples, then marginalise to obtain prediction.
- For a multiclass problem, i.e. $y \in \{C_1, \ldots, C_c\}$, we model the class conditional distribution $p(\mathbf{x}|y)$, and also the prior probabilities of each class $p(y)$. Then we compute the posterior probability for each class using

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

Discriminative approach

The discriminative approach models $p(y|\mathbf{x})$ directly.

Generative vs Discriminative (2)

- Note that both ways of writing $p(\mathbf{x}, y)$ are correct!
- Discriminative approach directly models what we want, i.e. $p(y|\mathbf{x})$.
- Density estimation for class conditional distribution $p(\mathbf{x}|y)$ is a hard problem.
- Using the generative model, we can marginalize out the class label y from the joint by $p(\mathbf{x}) = \sum_y p(y)p(\mathbf{x}|y)$. This gives us a distribution from which we can draw examples.
- Using the density, it is easy to deal with missing input values, outliers and unlabelled data in a principled fashion.

Examples

Generative

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

Discriminative

- SVM
- Maximum likelihood
- Maximum a Posteriori
- . . .

Maximum Likelihood

Linear regression

$$y = \mathbf{w}^\top \mathbf{x} + \varepsilon,$$

where ε is zero mean Gaussian noise.

Likelihood

The probability density of the observations given the parameters,

$$p(y|\mathbf{x}, \mathbf{w}),$$

is called the likelihood.

Maximum Likelihood

Maximizing $p(y|\mathbf{x}, \mathbf{w})$ with respect to $\mathbf{w}$ gives the parameters for which the probability of observed data is maximized.

Log Likelihood

Likelihood

The probability density of the observations given the parameters,

$$p(y|\mathbf{x}, \mathbf{w}),$$

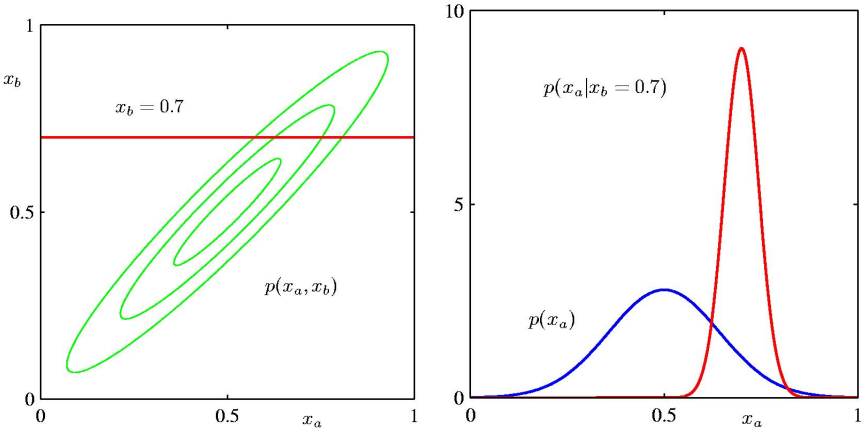
is called the likelihood.

Negative Log Likelihood

Usually minimize $-\log p(y|\mathbf{x}, \mathbf{w})$.

- Logarithm is monotonic
- Simplifies maths
- Better numerical behaviour

Gaussian Distribution (1)



1 dimensional

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

D dimensional

$$\mathcal{N}(\mathbf{x}|\mu,\mathbf{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right)$$

Gaussian Distribution (2)

D dimensional

$$\mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^\top \mathbf{\Sigma}^{-1}(\mathbf{x}-\mu)\right)$$

Negative Log Likelihood

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

Gaussian Noise

Linear Regression

$$y = \mathbf{w}^\top \mathbf{x} + \varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

Independence

Since $(\mathbf{x}_i, y_i)$ are independent,

$$p(y|\mathbf{x}, \mathbf{w}) = \prod_{i=1}^{n} p(y_i|\mathbf{x}_i, \mathbf{w})$$

Gaussian noise

Using the fact that ε is Gaussian, then $y_i \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}_i, \sigma^2)$, and hence

$$p(y|\mathbf{x}, \mathbf{w}) = \prod_{i=1}^{n} \mathcal{N}(y_i - \mathbf{w}^\top \mathbf{x}_i, \sigma^2)$$

Optimization

Maximum Likelihood with Gaussian Noise

$$p(y|\mathbf{x}, \mathbf{w}) = \prod_{i=1}^{n} \mathcal{N}(y_i - \mathbf{w}^\top \mathbf{x}_i, \sigma^2)$$

Negative Log Likelihood

$$\min_{\mathbf{w}} -\log p(y|\mathbf{x}, \mathbf{w}) = \sum_{i=1}^{n} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2$$

Least Squares

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

Solution

$$\mathbf{w}_{ML} = (X^\top X)^{-1} X^\top y$$

Probabilistic Model

Maximum Likelihood

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

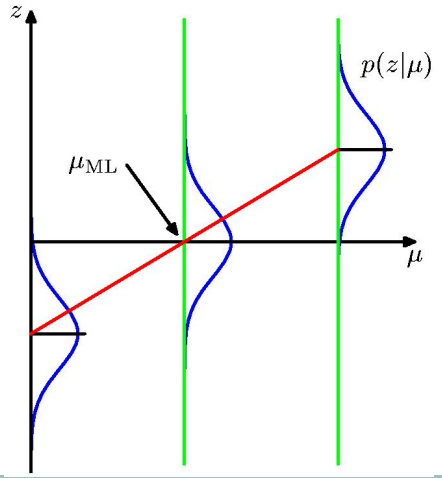
Recall

We have a probabilistic model, and hence we can make a predictive distribution for a new value $\mathbf{x}_t$.

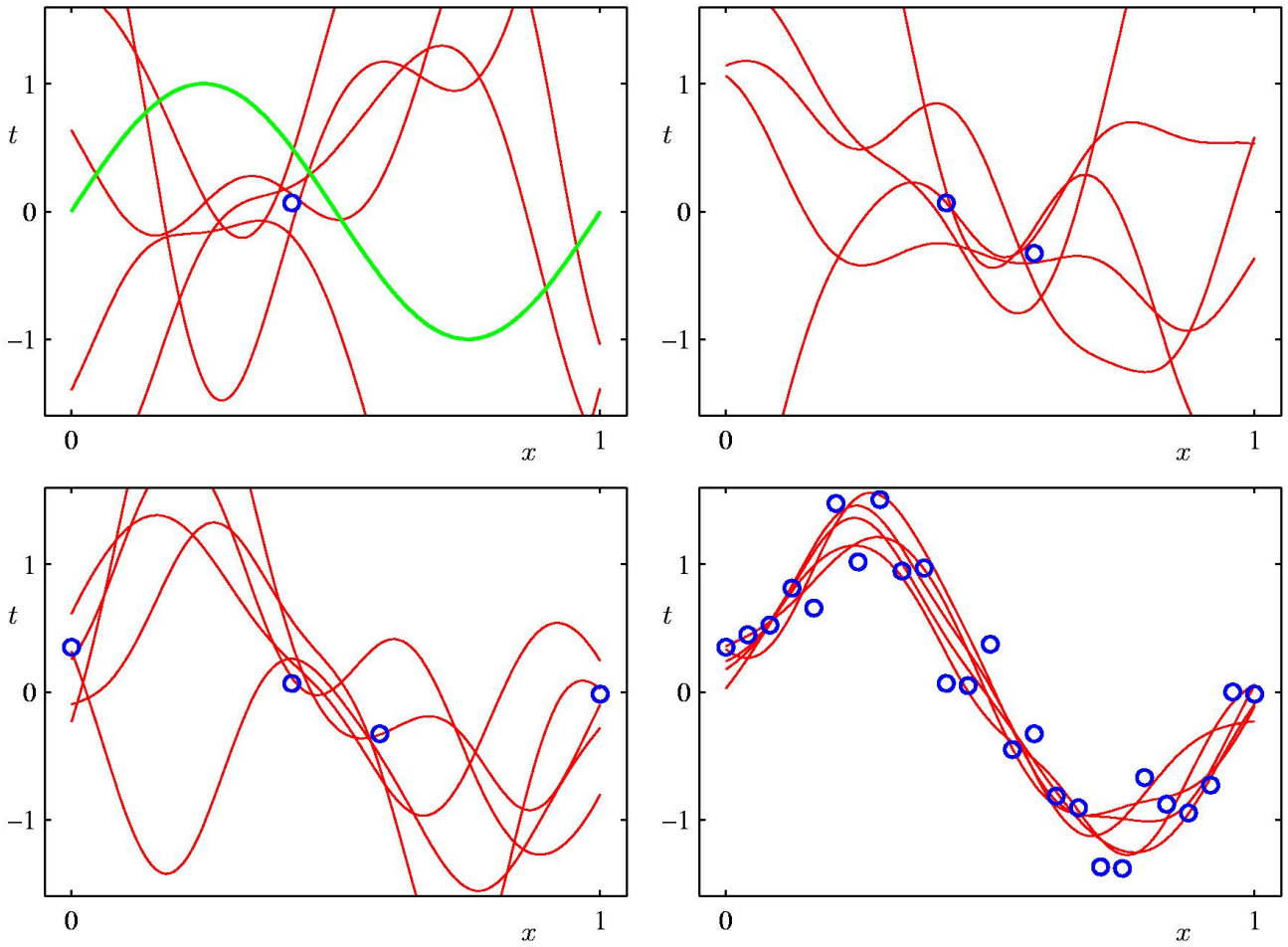
In other words

For each test example $\mathbf{x}_t$, we get a distribution

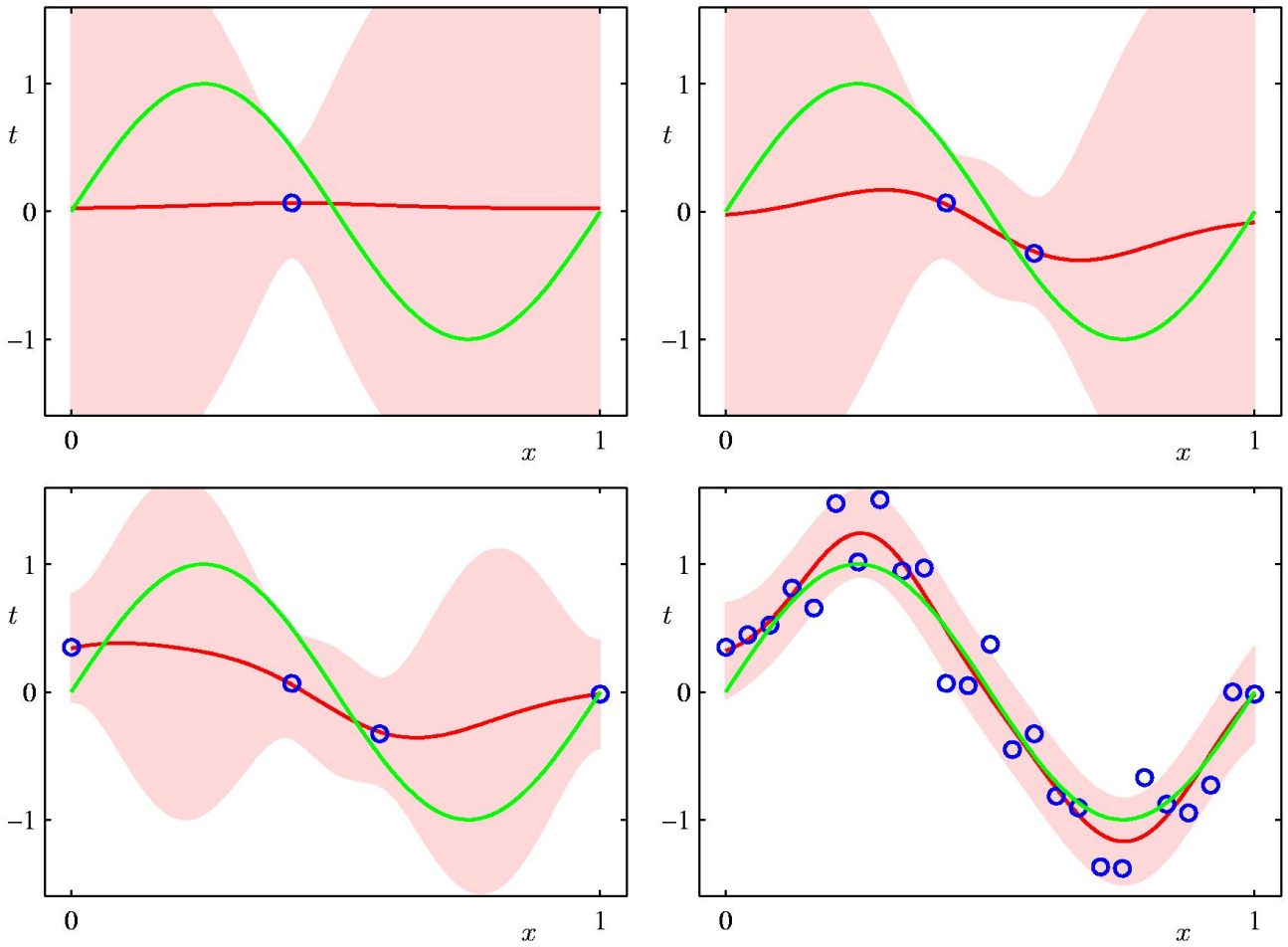
$$y_t \sim \mathcal{N}(\mathbf{w}_{ML}^\top \mathbf{x}_t, \sigma^2).$$



Regression: Samples



Regression: Variance



Prior

Bayes rule

$$p(\mathbf{w}|\mathbf{x}, y) = \frac{p(y|\mathbf{x}, \mathbf{w})p(\mathbf{w})}{p(y|\mathbf{x})}$$

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Using background information

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

$$\mathbf{w} \sim \mathcal{N}(0, \Sigma^{-1}).$$

i.e.

$$p(\mathbf{w}) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\mathbf{w}^\top \mathbf{\Sigma}^{-1}\mathbf{w}\right).$$

Maximum a Posteriori

Idea

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

Posterior

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

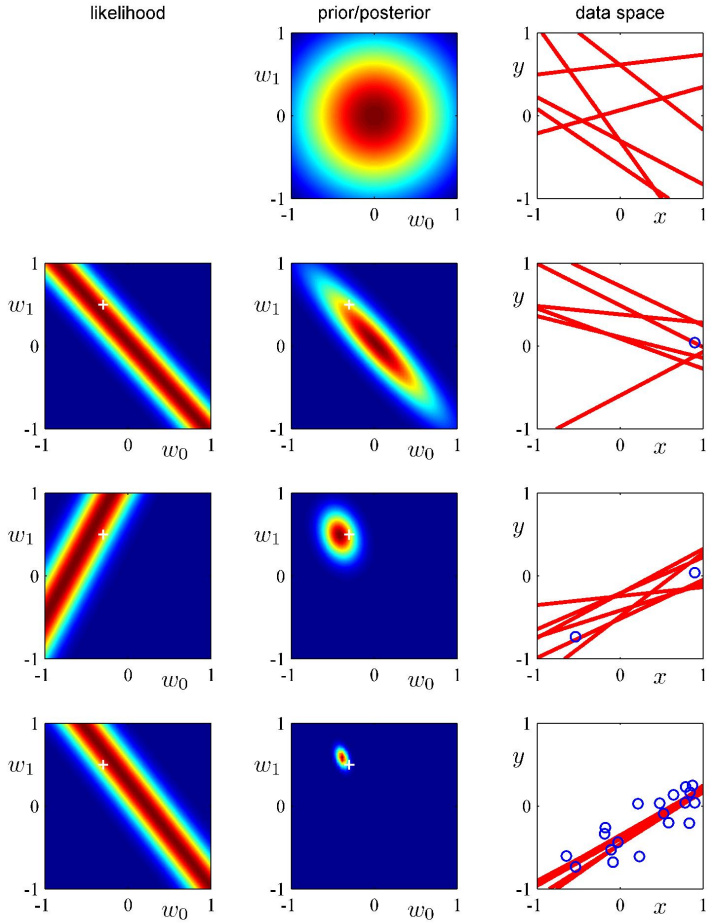
Negative Log Posterior

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

Regularized Least Squares

The negative log posterior is the regularized least squares problem.

Likelihood, Prior, Examples



Classification

Binary Classification

The class conditional densities are

$$p(\mathbf{x}|Y=1) \quad \text{and} \quad p(\mathbf{x}|Y=-1),$$

and the class priors are

$$p(Y=1) \quad \text{and} \quad p(Y=-1),$$

Recall

- $\frac{A}{A+B} = \frac{1}{1+\frac{B}{A}}$
- Bayes Theorem: $p(y|x) = \frac{p(x|y)p(y)}{p(x)}$
- Probabilities sum to one, specifically

$$p(\mathbf{x}) = p(\mathbf{x}|Y=1)p(Y=1) + p(\mathbf{x}|Y=-1)p(Y=-1).$$

Sigmoid Function

The posterior probability for the positive class is given by

$$p(Y=1|\mathbf{x}) = \frac{p(\mathbf{x}|Y=1)p(Y=1)}{p(\mathbf{x}|Y=1)p(Y=1)+p(\mathbf{x}|Y=-1)p(Y=-1)}$$
$$= \frac{1}{1+\exp(-a)}$$

where $a = \ln \frac{p(\mathbf{x}|Y=1)p(Y=1)}{p(\mathbf{x}|Y=-1)p(Y=-1)}$.

Sigmoid

$s(a) = \frac{1}{1+\exp(-a)}$ is called the sigmoid function ("S" shaped).

Log-Odds Ratio

What are the odds that the positive class is selected?

Gaussian again...

Assume Gaussian class conditional densities

$$p(\mathbf{x}|Y=1) = \mathcal{N}(\mu_+, \Sigma^{-1})$$

and

$$p(\mathbf{x}|Y=-1) = \mathcal{N}(\mu_-, \Sigma^{-1}).$$

From the results in the previous slide, the posterior is

$$p(Y=1|\mathbf{x}) = s(\mathbf{w}^\top x + w_0),$$

where

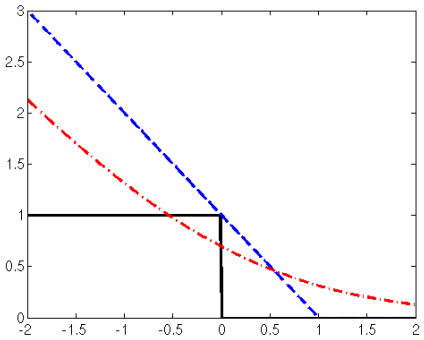
$$\mathbf{w} = \Sigma^{-1}(\mu_+ - \mu_-)$$

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

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

$$\ell(f(\mathbf{x}_i), y_i) := \log(1 + \exp(-y_i f(\mathbf{x}_i)))$$



- Can solve for the maximum likelihood (unregularized) or maximum a posteriori (regularized) weights.

Summary

- Probabilistic view provides different tools for analysis
- Bayes rule gives the relationship between the likelihood and the posterior
- From the probabilistic model, one can do both generative and discriminative methods.
- Maximum Likelihood with Gaussian noise is the least squares solution.
- Maximum a posteriori with a Gaussian prior is the regularized least squares solution.
- Logistic regression is a classification method, motivated by Gaussian class conditional densities.

References

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006.

R.O. Duda, P.E.Hart, and D.G.Stork. *Pattern classification*. John Wiley & Sons, second edition, 2001.

R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, 2002.

Jean Jacod and Philip Protter. *Probability Essentials*. Springer-Verlag, 2nd edition, 2004.

David J.C. Mackay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

Carl Edward Rasmussen and Christopher K.I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

A. N. Shiryayev. *Probability*. Springer-Verlag, 1984.