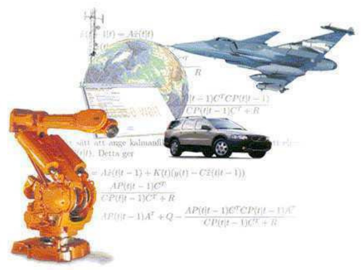
Machine Learning, Lecture 2
Linear Regression and Classification

"it is our firm belief that an understanding of linear models is essential for understanding nonlinear ones"



Thomas Schön

Division of Automatic Control
Linköping University
Linköping, Sweden.

Email: schon@isy.liu.se,
Phone: 013 - 281373,
Office: House B, Entrance 27.

Outline Lecture 2

1. Summary of Lecture 1
2. Bayesian linear regression
3. Motivation of Kernel methods
4. Linear Classification
    1. Problem setup
    2. Discriminant functions (mainly least squares)
    3. Probabilistic generative models
    4. Logistic regression (discriminative model)
    5. Bayesian logistic regression

(Chapter 3.3 - 4)

Summary of Lecture 1 (I/VI)

The exponential family of distributions over x, parameterized by η,

$$p(x \mid \eta) = h(x)g(\eta)\exp\left(\eta^T u(x)\right)$$

One important member is the Gaussian density, which is commonly used as a building block in more sophisticated models. Important basic properties were provided.

The idea underlying maximum likelihood is that the parameters θ should be chosen in such a way that the measurements $\{x_i\}_{i=1}^N$ are as likely as possible, i.e.,

$$\widehat{\theta} = \arg\max_{\theta} p(x_1, \cdots, x_N \mid \theta).$$

Summary of Lecture 1 (II/VI)

The three basic steps of Bayesian modeling (where all variables are modeled as stochastic)

1. Assign prior distributions $p(\theta)$ to all unknown parameters θ.
2. Write down the likelihood $p(x_1, \dots, x_N \mid \theta)$ of the data $x_1, \dots, x_N$ given the parameters θ.
3. Determine the posterior distribution of the parameters given the data

$$p(\theta \mid x_1, \dots, x_N) = \frac{p(x_1, \dots, x_N \mid \theta)p(\theta)}{p(x_1, \dots, x_N)} \propto p(x_1, \dots, x_N \mid \theta)p(\theta)$$

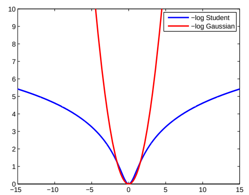
If the posterior $p(\theta \mid x_1, \dots, x_N)$ and the prior $p(\theta)$ distributions are of the same functional form they are conjugate distributions and the prior is said to be a conjugate prior for the likelihood.

Summary of Lecture 1 (III/VI)

Modeling "heavy tails" using the Student's t-distribution

$$\mathrm{St}(x \mid \mu, \lambda, \nu) = \int \mathcal{N}\left(x \mid \mu, (\eta\lambda)^{-1}\right) \mathrm{Gam}\,(\eta \mid \nu/2, \nu/2)\, d\eta$$
$$= \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)} \left(\frac{\lambda}{\pi\nu}\right)^{\frac{1}{2}} \left(1 + \frac{\lambda(x-\mu)^2}{\nu}\right)^{-\frac{\nu}{2}-\frac{1}{2}}$$

which according to the first expressions can be interpreted as an infinite mix of Gaussians with the same mean, but different variance.



Poor robustness is due to an unrealistic model, the ML estimator is inherently robust, provided we have the correct model.

Summary of Lecture 1 (IV/VI)

Linear regression models the relationship between a continuous target variable t and a possibly nonlinear function $\phi(x)$ of the input variable x,

$$t_n = \underbrace{w^T \phi(x_n)}_{y(x_n, w)} + \epsilon_n.$$

Solved this problem using

1. Maximum Likelihood (ML)
2. Bayesian approach

ML with a Gaussian noise model is equivalent to least squares (LS).

Summary of Lecture 1 (V/VI)

Theorem (Gauss-Markov)

In a linear regression model

$$T = \Phi w + E,$$

where E is white noise with zero mean and covariance R, the best linear unbiased estimate (BLUE) of w is

$$\widehat{w} = (\Phi^T R^{-1} \Phi)^{-1} \Phi^T R^{-1} T, \qquad \mathrm{Cov}(\widehat{w}) = (\Phi^T R^{-1} \Phi)^{-1}.$$

Interpretation: The least squares estimator has the smallest mean square error (MSE) of all linear estimators with no bias, BUT there may exist a biased estimator with lower MSE.

Summary of Lecture 1 (VI/VI)

Two potentially biased estimators are ridge regression ($p = 2$) and the Lasso ($p = 1$)

$$\begin{aligned} \min_w \quad & \textstyle\sum_{n=1}^N \left(t_n - w^T \phi(x_n)\right)^2 \\ \text{s.t.} \quad & \textstyle\sum_{j=0}^{M-1} |w_j|^p \leq \eta \end{aligned}$$

which using a Lagrange multiplier λ can be stated

$$\min_w \sum_{n=1}^N \left(t_n - w^T \phi(x_n)\right)^2 + \lambda \sum_{j=0}^{M-1} |w_j|^p$$

Alternative interpretation: The MAP estimate with the likelihood $\prod_{n=1}^N (t_n - w^T \phi(x_n))^2$ together with a Gaussian prior leads to ridge regression and together with a Laplacian prior leads to the LASSO.

Bayesian Linear Regression - Example (I/VI)

Consider the problem of fitting a straight line to noisy measurements. Let the model be ($t \in \mathcal{R}, x_n \in \mathcal{R}$)

$$t_n = \underbrace{w_0 + w_1 x_n}_{y(x,w)} + \epsilon_n, \qquad n = 1, \dots, N. \tag{1}$$

where

$$\epsilon_n \sim \mathcal{N}(0, 0.2^2), \qquad \beta = \frac{1}{0.2^2} = 25.$$

According to (1), the following identity basis function is used

$$\phi_0(x_n) = 1, \qquad \phi_1(x_n) = x_n.$$

Since the example lives in two dimensions, we can plot distributions to illustrate the inference.

Bayesian Linear Regression - Example (II/VI)

Let the true values for w be $w^\star = \begin{pmatrix} -0.3 & 0.5 \end{pmatrix}^T$ (plotted using a white circle below).

Generate synthetic measurements by

$$t_n = w_0^\star + w_1^\star x_n + \epsilon_n, \qquad \epsilon_n \sim \mathcal{N}(0, 0.2^2),$$

where $x_n \sim \mathcal{U}(-1, 1)$.

Furthermore, let the prior be

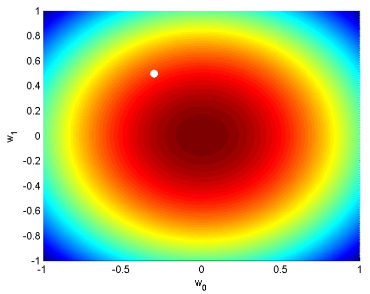
$$p(w) = \mathcal{N}\left(w \mid \begin{pmatrix} 0 & 0 \end{pmatrix}^T, \alpha^{-1} I\right)$$

where

$$\alpha = 2.$$

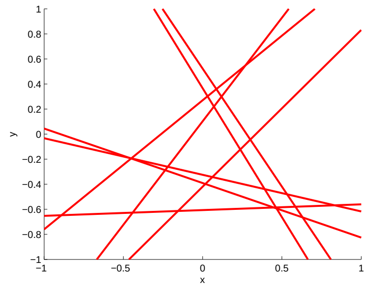
Bayesian Linear Regression - Example (III/VI)

Plot of the situation before any data arrives.



Prior,

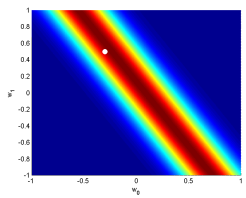
$$p(w) = \mathcal{N}\left(w \mid \begin{pmatrix} 0 & 0 \end{pmatrix}^T, \frac{1}{2} I\right)$$



Example of a few realizations from the posterior.

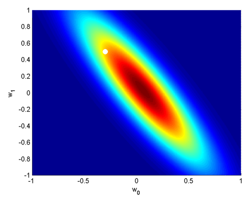
Bayesian Linear Regression - Example (IV/VI)

Plot of the situation after one measurement has arrived.



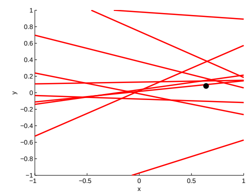
Likelihood (plotted as a function of w)

$$p(t_1 \mid w) = \mathcal{N}(t_1 \mid w_0 + w_1 x_1, \beta^{-1})$$



Posterior/prior,

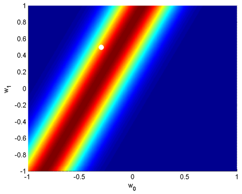
$$p(w \mid t_1) = \mathcal{N}(w \mid m_1, S_1),$$
$$m_1 = \beta S_1 \Phi^T t_1,$$
$$S_1 = (\alpha I + \beta \Phi^T \Phi)^{-1}.$$



Example of a few realizations from the posterior and the first measurement (black circle).

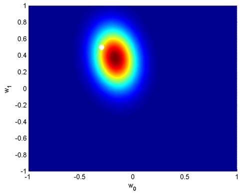
Bayesian Linear Regression - Example (V/VI)

Plot of the situation after two measurements have arrived.



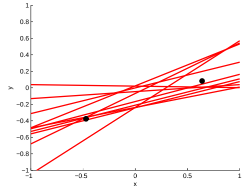
Likelihood (plotted as a function of w)

$$p(t_2 \mid w) = \mathcal{N}(t_2 \mid w_0 + w_1 x_2, \beta^{-1})$$



Posterior/prior,

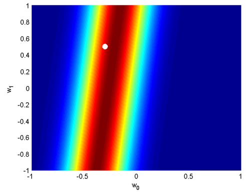
$$p(w \mid T) = \mathcal{N}(w \mid m_2, S_2),$$
$$m_2 = \beta S_2 \Phi^T T,$$
$$S_2 = (\alpha I + \beta \Phi^T \Phi)^{-1}.$$



Example of a few realizations from the posterior and the measurements (black circles).

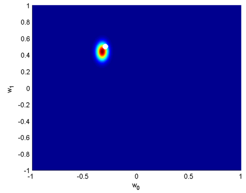
Bayesian Linear Regression - Example (VI/VI)

Plot of the situation after 30 measurements have arrived.



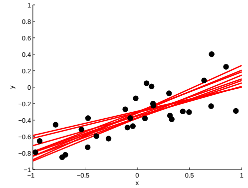
Likelihood (plotted as a function of w)

$$p(t_{30} \mid w) = \mathcal{N}(t_{30} \mid w_0 + w_1 x_{30}, \beta^{-1})$$



Posterior/prior,

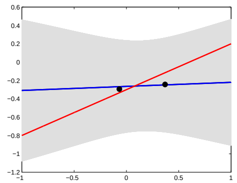
$$p(w \mid T) = \mathcal{N}(w \mid m_{30}, S_{30}),$$
$$m_{30} = \beta S_{30} \Phi^T T,$$
$$S_{30} = (\alpha I + \beta \Phi^T \Phi)^{-1}.$$



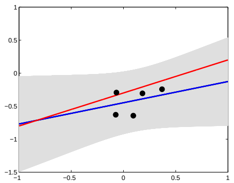
Example of a few realizations from the posterior and the measurements (black circles).

Predictive Distribution - Example

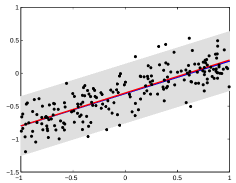
Investigating the predictive distribution for the example above



$N = 2$ observations



$N = 5$ observations



$N = 200$ observations

- True system ($y(x) = -0.3 + 0.5x$) generating the data (red line)
- Mean of the predictive distribution (blue line)
- One standard deviation of the predictive distribution (gray shaded area) Note that this is the *point-wise predictive standard deviation* as a function of x.
- Observations (black circles)

Posterior Distribution

Recall that the posterior distribution is given by

$$p(w \mid T) = \mathcal{N}(w \mid m_N, S_N),$$

where

$$m_N = \beta S_N \Phi^T T,$$
$$S_N = (\alpha I + \beta \Phi^T \Phi)^{-1}.$$

Let us now investigate the posterior mean solution m_N above, which has an interpretation that directly leads to the kernel methods (lecture 4), including Gaussian processes.

Generative and Discriminative Models

Approaches that model the distributions of both the inputs and the outputs are known as generative models. The reason for the name is the fact that using these models we can generate new samples in the input space.

Approaches that models the posterior probability directly are referred to as discriminative models.

ML for Probabilistic Generative Models (I/IV)

Consider the two class case, where the class-conditional densities $p(x \mid \mathcal{C}_k)$ are Gaussian and the training data is given by $\{x_n, t_n\}_{n=1}^N$. Furthermore, assume that $p(\mathcal{C}_1) = \alpha$.

The task is now to find the parameters $\alpha, \mu_1, \mu_2, \Sigma$ by maximizing the likelihood function,

$$p(T, X \mid \alpha, \mu_1, \mu_2, \Sigma) = \prod_{n=1}^{N} (p(x_n, \mathcal{C}_1))^{t_n} (p(x_n, \mathcal{C}_2)^{1-t_n},$$

where

$$\begin{aligned} p(x_n, \mathcal{C}_1) &= p(\mathcal{C}_1) p(x_n \mid \mathcal{C}_1) = \alpha \mathcal{N}(x_n \mid \mu_1, \Sigma), \\ p(x_n, \mathcal{C}_2) &= p(\mathcal{C}_2) p(x_n \mid \mathcal{C}_2) = (1-\alpha) \mathcal{N}(x_n \mid \mu_2, \Sigma). \end{aligned}$$

ML for Probabilistic Generative Models (II/IV)

Let us now maximize the logarithm of the likelihood function,

$$L(\alpha, \mu_1, \mu_2, \Sigma) = \ln \left(\prod_{n=1}^{N} (\alpha \mathcal{N}(x_n \mid \mu_1, \Sigma))^{t_n} ((1-\alpha) \mathcal{N}(x_n \mid \mu_2, \Sigma))^{1-t_n} \right)$$

The terms that depends on α are

$$\sum_{n=1}^{N} (t_n \ln \alpha + (1-t_n) \ln(1-\alpha))$$

which is maximized by $\widehat{\alpha} = \frac{1}{N} \sum_{n=1}^{N} t_n = \frac{N_1}{N_1 + N_2}$ (as expected). N_k denotes the number of data in class $\mathcal{C}_k$. Straightforwardly we get

$$\widehat{\mu}_1 = \frac{1}{N_1} \sum_{n=1}^{N} t_n x_n, \qquad \widehat{\mu}_2 = \frac{1}{N_2} \sum_{n=1}^{N} (1-t_n) x_n.$$

ML for Probabilistic Generative Models (III/IV)

$$\begin{aligned} L(\Sigma) = &-\frac{1}{2} \sum_{n=1}^{N} t_n \ln \det \Sigma - \frac{1}{2} \sum_{n=1}^{N} t_n (x_n - \mu_1)^T \Sigma^{-1} (x_n - \mu_1) \\ &- \frac{1}{2} \sum_{n=1}^{N} (1-t_n) \ln \det \Sigma - \frac{1}{2} \sum_{n=1}^{N} (1-t_n)(x_n - \mu_2)^T \Sigma^{-1} (x_n - \mu_2) \end{aligned}$$

Using the fact that $x^T A x = \operatorname{Tr}\left(A x x^T\right)$ we have

$$L(\Sigma) = -\frac{N}{2} \ln \det \Sigma - \frac{N}{2} \operatorname{Tr}\left(\Sigma^{-1} S\right),$$

where

$$S = \frac{1}{N} \sum_{n=1}^{N} \left(t_n (x_n - \mu_1)(x_n - \mu_1)^T + (1-t_n)(x_n - \mu_2)(x_n - \mu_2)^T \right)$$

ML for Probabilistic Generative Models (IV/IV)

Lemma (Useful Matrix derivatives)

$$\frac{\partial}{\partial M} \ln \det M = M^{-T},$$
$$\frac{\partial}{\partial M} \operatorname{Tr}\left(M^{-1}N\right) = -M^{-T}N^TM^{-T}.$$

Differentiating $L(\Sigma) = -\frac{N}{2}\ln\det\Sigma - \frac{N}{2}\operatorname{Tr}\left(\Sigma^{-1}S\right)$ results in

$$\frac{\partial L}{\partial \Sigma} = -\frac{N}{2}\Sigma^{-T} + \frac{N}{2}\Sigma^{-T}S\Sigma^{-T}$$
$$\frac{\partial L}{\partial \Sigma} = 0$$

Hence, $\Sigma = S$

More results on matrix derivatives are available in Magnus, J. R., & Neudecker, H. (1999). *Matrix differential calculus with applications in statistics and econometrics.* 2nd Edition, Chichester, UK: Wiley.

Generalized Linear Models for Classification

In linear regression we made use of a linear model

$$t_n = y(x, w) = w^T\phi(x_n) + \epsilon_n.$$

For classification problems the target variables are discrete, or slightly more general, posterior probabilities in the range $(0, 1)$. This is achieved using a so called *activation* function f (f^{-1} must exist),

$$y(x) = f(w^Tx + w_0). \tag{2}$$

Note that the decision surface corresponds to $y(x) =$ constant, implying that $w^Tx + w_0 =$ constant. This means that the decision surface is a linear function of x, even if f is nonlinear. Hence, the name *generalized linear model* for (2).

Gradient of $L(w)$ for Logistic Regression (I/II)

The negative log-likelihood is

$$L(w) = -\sum_{n=1}^{N}\left(t_n \ln y_n + (1-t_n)\ln(1-y_n)\right),$$

where

$$y_n = \sigma(a_n) = \frac{1}{1+\exp(-a_n)}, \qquad \text{and} \qquad a_n = w^T\phi_n.$$

Using the chain rule we have,

$$\frac{\partial L}{\partial w} = \sum_{n=1}^{N}\frac{\partial L}{\partial y_n}\frac{\partial y_n}{\partial a_n}\frac{\partial a_n}{\partial w}$$

where

$$\frac{\partial L}{\partial y_n} = \frac{1-t_n}{1-y_n} - \frac{t_n}{y_n} = \frac{y_n - t_n}{y_n(1-y_n)}$$

Gradient of $L(w)$ for Logistic Regression (II/II)

Furthermore,

$$\frac{\partial y_n}{\partial a_n} = \frac{\partial \sigma(a_n)}{\partial a_n} = \cdots = \sigma(a_n)(1-\sigma(a_n)) = y_n(1-y_n),$$
$$\frac{\partial a_n}{\partial w} = \phi_n.$$

which results in the following expression for the gradient

$$\frac{\partial L}{\partial w} = \sum_{n=1}^{N}(y_n - t_n)\phi_n = \Phi^T(Y - T),$$

where

$$\Phi = \begin{pmatrix} \phi_0(x_1) & \phi_1(x_1) & \dots & \phi_{M-1}(x_1) \\ \phi_0(x_2) & \phi_1(x_2) & \dots & \phi_{M-1}(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(x_N) & \phi_1(x_N) & \dots & \phi_{M-1}(x_N) \end{pmatrix} \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} \quad T = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{pmatrix}$$

Hessian of $L(w)$ for Logistic Regression

$$H = \frac{\partial^2 L}{\partial w \partial w^T} = \cdots = \sum_{n=1}^{N} (y_n - t_n)\phi_n \phi_n^T = \Phi^T R \Phi$$

where

$$R = \begin{pmatrix} y_1(1-y_1) & 0 & \dots & 0 \\ 0 & y_2(1-y_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & y_N(1-y_N) \end{pmatrix}$$

Bayesian Logistic Regression

Recall that

$$p(T \mid w) = \prod_{n=1}^{N} \sigma(w^T \phi_n)^{t_n} \left(1 - \sigma(w^T \phi_n)\right)^{1-t_n}$$

Hence, computing the posterior density

$$p(w \mid T) = \frac{p(T \mid w) p(w)}{p(T)}$$

is intractable. We are forced to an approximation. Three alternatives

1. Laplace approximation (this lecture)
2. VB & EP (lecture 5)
3. Sampling methods, e.g., MCMC (lecture 6)

Laplace Approximation (I/III)

The Laplace approximation is a simple approximation that is obtained by fitting a Gaussian centered around the (MAP) mode of the distribution.

Consider the density function $p(z)$ of a scalar stochastic variable z, given by

$$p(z) = \frac{1}{Z} f(z),$$

where $Z = \int f(z) dz$ is the normalization coefficient.

We start by finding a mode z_0 of the density function,

$$\left.\frac{df(z)}{dz}\right|_{z=z_0} = 0.$$

Laplace Approximation (II/III)

Consider a Taylor expansion of $\ln f(z)$ around the mode z_0,

$$\begin{aligned} \ln f(z) &\approx \ln f(z_0) + \underbrace{\left.\frac{d}{dz} \ln f(z)\right|_{z=z_0}}_{=0} (z - z_0) + \frac{1}{2} \left.\frac{d^2}{dz^2} \ln f(z)\right|_{z=z_0} (z - z_0)^2 \\ &= \ln f(z_0) - \frac{A}{2}(z - z_0)^2, \qquad (3) \end{aligned}$$

where

$$A = -\left.\frac{d^2}{dz^2} \ln f(z)\right|_{z=z_0}$$

Taking the exponential of both sides in the approx. (3) results in

$$f(z) \approx f(z_0) \exp\left(-\frac{A}{2}(z - z_0)^2\right)$$

Laplace Approximation (III/III)

By normalizing this expression we have now obtained a Gaussian approximation

$$q(z) = \left(\frac{A}{2\pi}\right)^{1/2} \exp\left(-\frac{A}{2}(z - z_0)^2\right)$$

where

$$A = -\frac{d^2}{dz^2} \ln f(z)\Big|_{z=z_0}$$

The main limitation of the Laplace approximation is that it is a local method that only captures aspects of the true density around a specific value z_0.

Bayesian Logistic Regression (I/II)

The posterior is

$$p(w \mid T) \propto p(T \mid w)p(w), \tag{4}$$

where we assume a Gaussian prior $p(w) = \mathcal{N}(w \mid m_0, S_0)$ and make use of the Laplace approximation. Taking logarithm on both sides of (4) gives

$$\ln p(w \mid t) = -\frac{1}{2}(w - m_0)^T S_0^{-1}(w - m_0) + \sum_{n=1}^{N} (t_n \ln y_n + (1 - t_n) \ln(1 - y_n)) + \text{const.}$$

where $y_n = \sigma(w^T \phi_n)$.

Bayesian Logistic Regression (II/II)

Using the Laplace approximation we can now obtain a Gaussian approximation

$$q(w) = \mathcal{N}(w \mid w_{\mathsf{MAP}}, S_N)$$

where w_{MAP} is the MAP estimate of $p(w \mid T)$ and the covariance S_N is the Hessian of $\ln p(w \mid T)$,

$$S_N = \frac{\partial^2}{\partial w \partial w^T} \ln p(w \mid T) = S_0^{-1} + \sum_{n=1}^{N} y_n(1 - y_n)\phi_n \phi_n^T$$

Based on this distribution we can now start making predictions for new input data $\phi(x)$, which is typically what we are interested in. Recall that prediction corresponds to marginalization w.r.t. w.

A Few Concepts to Summarize Lecture 2

Hyperparameter: A parameter of the prior distribution that controls the distribution of the parameters of the model.
Classification: The goal of classification is to assign an input vector x to one of K classes, $\mathcal{C}_k, k = 1, \ldots, K$.

Discriminant: A discriminant is a function that takes an input x and assigns it to one of K classes.

Generative models: Approaches that model the distributions of both the inputs and the outputs are known as generative models. In classification this amounts to modelling the class-conditional densities $p(x \mid \mathcal{C}_k)$, as well as the prior densities $p(\mathcal{C}_k)$. The reason for the name is the fact that using these models we can generate new samples in the input space.

Discriminative models: Approaches that models the posterior probability directly are referred to as discriminative models.

Logistic Regression: Discriminative model that makes direct use of a generalized linear model in the form of a logistic sigmoid to solve the classification problem.

Laplace approximation: A local approximation method that finds the mode of the posterior distribution and then fits a Gaussian centered at that mode.