

Machine Learning, Lecture 3 Linear Classification



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 3

2(21)

1. Summary of lecture 2
2. Problem setup
3. Discriminant functions (mainly least squares)
4. Probabilistic generative models
5. Logistic regression (discriminative model)
6. Bayesian logistic regression



Summary of lecture 2 (I/III)

3(21)

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 2 (II/III)

4(21)

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

$$\hat{w} = (\Phi^T R^{-1} \Phi)^{-1} \Phi^T R^{-1} T, \quad \text{Cov}(\hat{w}) = (\Phi^T R^{-1} \Phi)^{-1}.$$

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.



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

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

which using a Lagrange multiplier λ can be stated

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

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 it leads to the LASSO.

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**.

Consider the two class case, where the class-conditional densities $p(x | \mathcal{C}_k)$ 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 (i.i.d. Bernoulli),

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

where

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

Let us now maximize the logarithm of the likelihood function,

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

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

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

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

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

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

where

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

Lemma (Useful matrix derivatives)

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

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

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

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, 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),

$$y(x) = f(w^T x + w_0). \quad (1)$$

Note that the decision surface corresponds to $y(x) = \text{constant}$, implying that $w^T x + w_0 = \text{constant}$. This means that the decision surface is a linear function of x , even if f is nonlinear. Hence, the name *generalized linear model* for (1).

Gradient of $L(w)$ for logistic regression (I/II)

The negative log-likelihood is

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

where

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

Using the chain rule we have,

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

where

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

Furthermore,

$$\frac{\partial y_n}{\partial a_n} = \frac{\partial \sigma(a_n)}{\partial a_n} = \dots = \sigma(a_n)(1 - \sigma(a_n)) = y_n(1 - y_n),$$

$$\frac{\partial a_n}{\partial w} = \phi_n.$$

which results in the following expression for the gradient

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

where

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

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

where

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

Recall that

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

Hence, computing the posterior density

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

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

1. Laplace approximation (this lecture)
2. VB & EP (lecture 7)
3. Sampling methods, e.g., MCMC (lecture 10)

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.

1. Find the MAP mode: We start by finding a mode z_0 of the density function,

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

2. Compute a Gaussian approximation around the MAP mode:

Start by considering a Taylor expansion of $\ln f(z)$ around z_0 ,

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

where

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

Taking the exponential of both sides in (2) results in

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

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 .

The posterior is

$$p(w | T) \propto p(T | w) p(w), \quad (3)$$

where we assume a Gaussian prior $p(w) = \mathcal{N}(w | m_0, S_0)$ and make use of the Laplace approximation. Taking the logarithm of both sides of (3) gives

$$\begin{aligned} \ln p(w | T) &= - \frac{1}{2} (w - m_0)^T S_0^{-1} (w - m_0) \\ &\quad + \sum_{n=1}^N (t_n \ln y_n + (1 - t_n) \ln(1 - y_n)) + \text{const.} \end{aligned}$$

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

Using the Laplace approximation we can now obtain a Gaussian approximation

$$q(w) = \mathcal{N}(w | w_{\text{MAP}}, S_N)$$

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

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

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 .

Classification: The goal of classification is to assign an input vector x to one of K classes, $C_k, k = 1, \dots, K$.

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 $p(x | C_k)$, as well as the prior densities $p(C_k)$. 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.

