Statistical Machine Learning

Lecture 3 - Linear classification

Thomas Schön
Division of Systems and Control
Department of Information Technology
Uppsala University.

Email: thomas.schon@it.uu.se,
www: `user.it.uu.se/~thosc112`

UPPSALA UNIVERSITET

UPPSALA UNIVERSITET

Contents – lecture 3

UPPSALA UNIVERSITET

Summary of lecture 2 (I/III)

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

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

We solved this problem using:

1. Maximum Likelihood (ML)
2. Bayesian approach

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

UPPSALA UNIVERSITET

Summary of lecture 2 (II/III)

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

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

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 2 (III/III)

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

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

which using a Lagrange multiplier λ can be stated

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

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

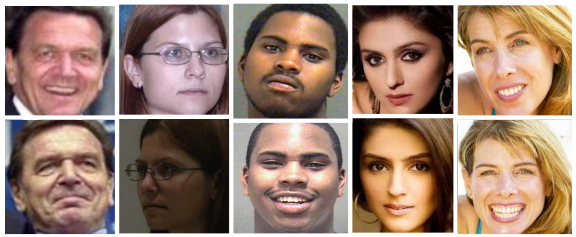
Generative and discriminative models

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**.

A recent classification example

Aim: Determine whether two face images are from the same person or not (a problem known as "face verification" in CV).



Lu, C. and Tang, X. **Surpassing Human-Level Face Verification Performance on LFW with GaussianFace**. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI)*, Austin, TX, USA, 2015.

A recent classification example

Key elements of their solution:

- Flexible Gaussian Process (GP) model. (Recall the "alternative approach" to linear regression from last lecture, see also Lecture 5). Latent variable are key here.
- Kernel Fisher discriminant classifier using the GP kernel.
- (The **Laplace approx.** is used to approx. the posterior.)

The resulting GP model can be used in two ways:

1. As a binary classifier
2. As a feature extractor

Human performance: 97.5%
GaussianFace performance: 98.5%
There are now even better solutions based on deep learning...

ML for probabilistic generative 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

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

ML for probabilistic generative models

Let us now maximize the logarithm of the likelihood function,

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

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 $\mathcal{C}_k$. Furthermore,

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

ML for probabilistic generative models

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

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

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

where

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

ML for probabilistic generative models

Lemma (Useful matrix derivatives)

$$\begin{aligned} \frac{\partial}{\partial M} \ln \det M &= M^{-\mathsf{T}}, \\ \frac{\partial}{\partial M} \operatorname{Tr}\left(M^{-1} N\right) &= -M^{-\mathsf{T}} N^\mathsf{T} M^{-\mathsf{T}}. \end{aligned}$$

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

$$\begin{aligned} \frac{\partial L}{\partial \Sigma} &= -\frac{N}{2} \Sigma^{-\mathsf{T}} + \frac{N}{2} \Sigma^{-\mathsf{T}} S \Sigma^{-\mathsf{T}} \\ \frac{\partial L}{\partial \Sigma} &= 0 \end{aligned}$$

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, UK: Wiley.

Generalized linear models for classification

In linear regression we made use of a linear model

$$t_n = y(x, w) = w^\mathsf{T}\phi(x_n) + \varepsilon_n.$$

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),

$$y(x, w) = f(w^\mathsf{T}x + w_0). \tag{1}$$

Note that the decision surface corresponds to $y(x, w) =$ constant, implying that $w^\mathsf{T}x + w_0 =$ 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 (1).

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

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

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

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

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

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

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

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

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

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

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 7)
3. Sampling methods, e.g., MCMC (lecture 10)

Laplace approximation (I/III)

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

Consider the density function $p(z)$ of a scalar random variable z,

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

where $Z = \int f(z)\mathrm{d}z$ is the normalization coefficient.

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

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

Laplace approximation (II/III)

2. Compute a Gaussian approximation around the mode: Start by considering a Taylor expansion of $\ln f(z)$ around z_0,

$$\ln f(z) \approx \ln f(z_0) + \underbrace{\left.\frac{d}{dz}\ln f(z)\right|_{z=z_0}}_{=0}(z - z_0) + \frac{1}{2}\frac{d^2}{dz^2}\ln f(z)\Big|_{z=z_0}(z -$$

$$= \ln f(z_0) - \frac{A}{2}(z - z_0)^2, \qquad (2)$$

where

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

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

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

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

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

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

$$p(w \,|\, T) \propto p(T \,|\, w) p(w), \tag{3}$$

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

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

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

Bayesian logistic regression (II/II)

Using the Laplace approximation we can now obtain a Gaussian approximation

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

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

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

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 3

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 $p(x \,|\, \mathcal{C}_k)$, 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.