CS281A/Stat241A: Statistical Learning Theory

Lecture 11: Linear Regression and Classification

Lecturer: Michael I. Jordan *Scribes: Tao Shi and Manikandan Narayanan*

1 Introduction

Today's strategy will be to race through Chapter 7 on "Linear Classification", covering important material briefly and leaving the rest for self-reading. We will focus on a simple conditional graphical model, where there is a single node conditioned on its multiple parents (see Figure 1). The model shown in the figure implies I.I.D. assumptions about the set of observed X_ns (where each X_n is made of p components). Richer models will be looked at later in the course.

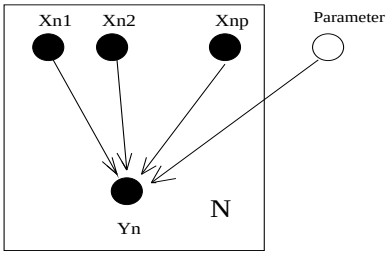


Figure 1: A simple conditional model

2 Linear Regression

In Linear Regression, we model the response of a continuous output variable Y_n to changes in the covaraiate x_n as

$Y_n = \beta^T x_n + \epsilon_n,$

where β is a parameter vector and ϵ_n is a random variable having zero mean. If $\epsilon_n \sim N(0, \sigma^2)$, then it follows that $P(y_n|x_n, \theta)$ is also normally distributed, with parameters depending on X_n and θ (where $\theta \triangleq (\beta, \sigma^2)$). That is,

$$Y_n|x_n, \theta \sim N(\beta^T x_n, \sigma^2).$$

Now, it is easy to compute the ML estimates of β and σ^2 by considering the log-likelihood. We sketch the calculations of $\hat{\beta}_{ML}$ here and leave those of $\hat{\sigma^2}_{ML}$ as an exercise.

$$\begin{aligned} l(\beta) &= -\frac{1}{2}\sum_{n=1}^{N}(y_n - \beta^T x_n)^2 \\ &= -\frac{1}{2}(Y - X\beta)^T(Y - X\beta) \end{aligned}$$

X is the design matrix, whose rows are the observations of the different x_n and whose columns correspond to the components of x_n. Y is just a vector of all y_ns stacked together. Setting $\bigtriangledown_\beta l = X^T(Y - X\beta) = 0$, we get the normal equation:

$$X^T X\hat{\beta}_{ML} = X^T Y$$

Solving the normal equation gives the ML estimate of β. In practice, instead of inverting $X^T X$, we use other methods like Gaussian elimination for speed and stability purposes. If $X^T X$ is singular, then SVD (Singular Value Decomposition) methods can be used.

An important remark should be made here. The model is linear in β, and not in X. If we map the data X to some feature space $f(X)$ nonlinear of X, we get $y_n = \beta^T f(x_n) + \epsilon_n$. It is not linear in X any longer, but it is still linear in β permitting the usage of linear regression machinery.

EXAMPLE 1: Let $f(x_n) = (1,\ x_n,\ x_n^2)^T$, we get polynomial regression in order 2:

$$y_n = \beta_1 + \beta_2 x_n + \beta_3 x_n^2 + \epsilon_n$$

3 Linear classification

In the classification case, the response Y_n is a discrete variable that depends on the covariate feature vector X_n. For example, we might be interested in classifying red and white wines, depending on their different "features" like sweetness, acidity, etc., In this case, $y_n \in \{0, 1\}$, where 0 represents the class 0 of red wines and 1 represents the class 1 of white wines. Note that it is not a good idea to treat y_n as continuous variable and do classification by linear regression. This is because, any data points that are outliers don't have much to do with the classification boundary. But they can have a strong influence on the parameter estimation (through the L^2 loss function) and change the classification boundary in undesired ways (see Figure 2).

We introduce some terminology and models related to the Classification problem here. We search for linear classifier in the space spanned by X and by Bayes rule, we have:

$$P(Y_n = 1|x_n) = \frac{P(x_n|Y_n = 1)P(Y_n = 1)}{P(x_n)}$$

with each term named as:

$P(Y_n = 1|x_n)$ — Posterior Distribution;
$P(x_n|Y_n = 1)$ — Class Conditional Distribution;
$P(Y_n = 1)$ — Prior Distribution.

There are two different ways to model this classification problem:

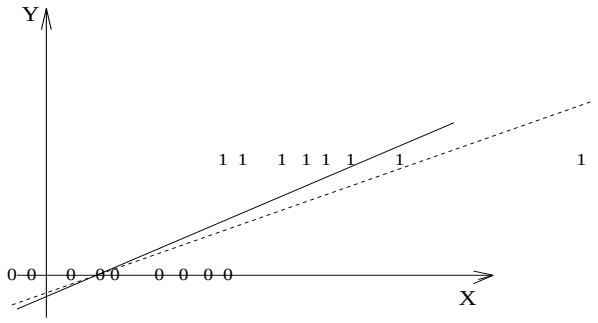


Figure 2: Linear regression in classification

- Generative model: Model two distributions - $P(x_n|Y_n = i)$ and $P(Y_n = i)$.
- Discriminative model: Model a single distribution - $P(Y_n = i|x_n)$.

From our point of view, they are just graphical models with arrows reversed (see Figure 3). In a closely related problem called the prediction problem, we know x_{new} but y_{new} is unknown. Our task is then to predict y_{new}, which can be done using either generative or discriminative models as shown in Figure 3.

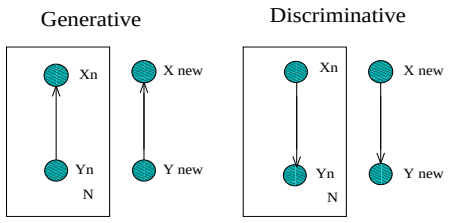


Figure 3: Generative model v.s. Discriminative model

4 Generative Model Examples

4.1 Gaussian class-conditionals

We will work through a binary classification example, where the class-conditional densities are normally distributed. The class label Y takes on one of the two values 0 or 1 (representing class 0 and class 1), with probability π and $1-\pi$ respectively. The class-conditional density of the p-dimensional feature vector X, is a multivariate Gaussian distribution on its features. So, we have the prior and the class-conditional for class k given by,

$$\pi \triangleq p(Y=1)$$
$$p(x \mid Y=k,\theta) = \frac{1}{(2\pi)^{p/2} \mid \Sigma \mid^{1/2}} \exp\{-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)\}.$$

The mean vectors of the two classes are μ_0 and μ_1, while the covariance matrix Σ, is assumed to be the same for both classes. All parameters are represented by $\theta \triangleq (\pi, \mu_0, \mu_1, \Sigma)$. Now, we compute the posterior using Bayes rule as follows:

$$\begin{aligned}
p(Y=1 \mid x,\theta) &= \frac{p(x \mid Y=1,\theta)p(Y=1 \mid \theta)}{p(x \mid Y=1,\theta)p(Y=1 \mid \theta) + p(x \mid Y=0,\theta)p(Y=0 \mid \theta)} \\
&= \frac{1}{1+\frac{p(x|Y=0,\theta)p(Y=0|\theta)}{p(x|Y=1,\theta)p(Y=1|\theta)}} \\
&= \frac{1}{1+\frac{1-\pi}{\pi}\exp\{\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - \frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\}} \\
&= \frac{1}{1+\exp\{-(\mu_1-\mu_0)^T\Sigma^{-1}x + \frac{1}{2}(\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1+\mu_0) - \log\frac{\pi}{1-\pi}\}} \\
&= \frac{1}{1+\exp\{-\beta^T x - \gamma\}} \\
&= \frac{1}{1+e^{-z}},
\end{aligned}$$

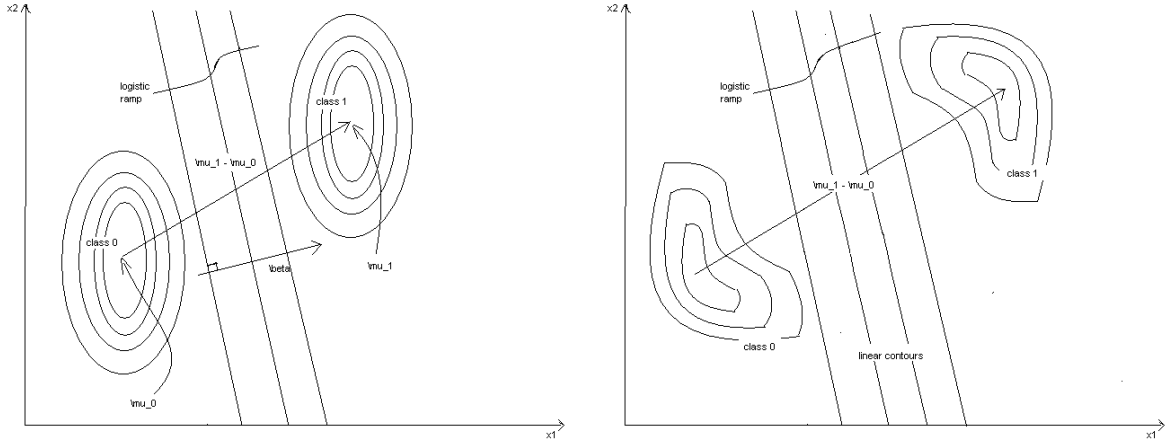
where $\beta \triangleq \Sigma^{-1}(\mu_1-\mu_0)$, $\gamma \triangleq -\frac{1}{2}(\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1+\mu_0) + \log\frac{\pi}{1-\pi}$ and $z \triangleq \beta^T x + \gamma$.

The posterior is the well-known logistic function, seen many times in this course. Further, the contours of equal posterior probability will be lines in the X-space, because z is an affine function of x. Thus, we get a linear classifier with the contour lines orthogonal to β, as shown in Figure 5(a). Note that in the figure, the contours of the two class-conditionals are centered around μ_0 and μ_1, and the posterior probability takes the form of a logistic ramp between the two classes. If we had assumed different Σ for the two classes, the posterior will still take the form of a logistic ramp between the two classes. But, it will now be a logistic of a quadratic (i.e., z would be a quadratic in x) and the contours of equal posterior probability will be quadratic boundaries in the X-space.

4.2 Other class-conditionals

We saw that Gaussian class-conditionals lead to logistic functions for posterior distributions. A natural question is: Does a logistic posterior imply a Gaussian class-conditional? Intuitively, the answer is no, because Figure 5(b) shows a non-Gaussian class-conditional having lines as contours of equal posterior probability (suggesting a logistic ramp function between the two classes). Infact, we will show here that there are distributions in the exponential family, if chosen for class conditionals will lead to a logistic posterior

Figure 4: Geometric Interpretation of Binary Classification



(a) Gaussian class-conditional on two features

(b) Non-Gaussian class-conditionals on two features

function. As we saw in the homework, exponential family distributions cover a wide range of continuous and discrete distributions and they have probability of the form, $p(x \mid \eta) = h(x) \exp\{\eta^T T(x) - A(\eta)\}$. If the class-conditionals are chosen from the exponential family with $T(x) = x$, then for the two classes,

$$p(x \mid Y = k, \theta) = h(x) \exp\{\eta_k^T x - A(\eta_k)\} \text{ and}$$
$$p(Y = 1 \mid x, \theta) = \frac{1}{1 + \exp\{-(\eta_0 - \eta_1)^T x - A(\eta_0) + A(\eta_1)\}}$$

which is a logistic of a function in x. Therefore, exponential family class-conditionals with $T(x) = x$ lead to posteriors, which are logistic of a linear function. The reverse direction is not true though, because there exists some non-trivial examples, where a logistic linear posterior results from a class-conditional not in the exponential family.