Linear Classifiers

김지수 (Jisu KIM)

통계적 기계학습(Statistical Machine Learning), 2024 1st semester

The lecture note is a minor modification of the lecture notes from Prof. Yongdai Kim's "Statistical Machine Learning".Also, see Section 4 from [1].

1 Review

1.1 Basic Model for Supervised Learning

- Input(입력) / Covariate(설명 변수) : $x \in \mathbb{R}^p$, so $x = (x_1, \ldots, x_p)$.
- Output(출력) / Response(반응 변수): $y \in \mathcal{Y}$. If y is categorical, then supervised learning is "classification", and if y is continuous, then supervised learning is "regression".
- Model(모형) :

  $$y \approx f(x).$$

  If we include the error ϵ to the model, then it can be also written as

  $$y = \phi(f(x), \epsilon).$$

  For many cases, we assume additive noise, so

  $$y = f(x) + \epsilon.$$

- Assumption(가정): f belongs to a family of functions $\mathcal{F}$. This is the assumption of a model: a model can be still used when the corresponding assumption is not satisfied in your data.
- Loss function(손실 함수): $\ell(y, a)$. A loss function measures the difference between estimated and true values for an instance of data.
- Training data(학습 자료): $\mathcal{T} = \{(y_i, x_i), i = 1, \ldots, n\}$, where (y_i, x_i) is a sample from a probability distribution P_i. For many cases we assume i.i.d., or x_i's are fixed and y_i's are i.i.d..
- Goal(목적): we want to find f that minimizes the expected prediction error,

  $$f^0 = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(Y,X)\sim P} \left[\ell(Y, f(X))\right].$$

- Prediction model(예측 모형): f^0 is unknown, so we estimate f^0 by $\hat{f}$ using data. For many cases we minimizes on the empirical prediction error, that is taking the expectation on the empirical distribution $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{(Y_i,X_i)}$.

  $$\hat{f} = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{P_n} \left[\ell(Y, f(X))\right] = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(X_i)).$$

- Prediction(예측): if $\hat{f}$ is a predicted function, and x is a new input, then we predict unknown y by $\hat{f}(x)$.

1.2 Bayes Classifier

For the classification so that $y \in \mathcal{Y} = \{1, \ldots, J\}$, the Expected Prediction Error (EPE) is

$$\begin{aligned} EPE(f) &= \mathbb{E}_{(Y,X)\sim P}\left[\ell(Y, f(X))\right] = \mathbb{E}\left[I(Y \neq f(X))\right] \\ &= \mathbb{E}_X \left[\sum_{j=1}^{J} \ell(j, f(X)) P(Y = j|X)\right]. \end{aligned}$$

Hence $EPE(f)$ is minimized when

$$f^0(x) = \arg\min_{k=1,\ldots,J} \sum_{j=1}^{J} \ell(j,k) P(Y = j|X = x).$$

For the classification, the most common loss is the 0-1 loss

$$\ell(y, a) = I(y \neq a).$$

And the optimal prediction function is

$$f^0(x) = \arg\max_{j=1,\ldots,J} P(Y = j|X = x).$$

This optimal classifier is called Bayes rule / Bayes classifier (베이즈분류 / 베이즈모형), and the error rate is called the Bayes risk (베이즈위험).

2 Linear Classifiers

We consider the K-class classification problem as

$$\mathcal{Y} = \{1, 2, \ldots, K\}.$$

The loss function for classification is the 0-1 loss given as

$$l(y, a) = I(y \neq a).$$

Let the classifier(분류기) $G : \mathbb{R}^p \to \mathcal{Y}$ be returning an output label given input. A decision boundary (결정 경계) is the region of a problem space in which the output label of a classifier is ambiguous. In other words, A decision boundary between class i and j is the set of points whose neighbor always intersect both regions classified as label i and label j, i.e.,

$$\{x \in \mathbb{R}^p : \text{for all } r > 0,\ B(x,r) \cap G^{-1}\{i\} \neq \emptyset,\ B(x,r) \cap G^{-1}\{j\} \neq \emptyset\},$$

where $B(x,r) = \{y \in \mathbb{R}^p : \|y - x\|_2 < r\}$ is the ball centered at x and radius r. And the decision boundary is the union over all pairs i and j, i.e.,

$$\bigcup_{1 \leq i < j \leq K} \{x \in \mathbb{R}^p : \text{for all } r > 0,\ B(x,r) \cap G^{-1}\{i\} \neq \emptyset,\ B(x,r) \cap G^{-1}\{j\} \neq \emptyset\}.$$

Linear methods for classification assume that the decision boundary between class i and j is given as a subset of an affine hyperspace

$$\{x : \beta_0 + x^\top \beta = 0\}.$$

There are three approaches for linear classifier(선형 분류기):

- model joint probability distribution (결합확률분포) : Linear Discriminant Analysis (선형판별분석), Naive Bayes Classifier (나이브 베이즈 분류기)
- model conditional density function $P(\text{class}|X)$: many classifier (분류기) based on regressions (회귀분석), in particular, Logistic Classification (로지스틱 분류) using Logistic Regression (로지스틱 회귀)
- Maximize difference between differeng groups: perceptron (퍼셉트론), support vector machine (서포트 벡터 머신)

This class covers Linear Discriminant Analysis (선형판별분석) and Logistic Classification (로지스틱 분류).

3 Logistic Classification (로지스틱 분류)

The logistic model assumes that for $1 \le k \le K-1$,

$$\Pr(y=k|x) = \frac{\exp(\beta_{k0} + x^\top \beta_k)}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + x^\top \beta_l)},$$

and

$$\Pr(y=K|x) = \frac{1}{1 + \sum_{l=1}^{K-1} \exp(\beta_{l0} + x^\top \beta_l)}.$$

In other words, $\beta_{K0} = 0$ and $\beta_K = 0$, which is for identifiablity problem. The Bayes classifier is

$$G(x) = \arg \max_{j=1,\dots,K} P(y=j|x).$$

Now, the decision boundary for i and j is a subset of where $\Pr(y=i|x) = \Pr(y=j|x)$, and this is a hyperplane defined as

$$\left\{ x \in \mathbb{R}^p : (\beta_{i0} - \beta_{j0}) + x^\top (\beta_i - \beta_j) = 0 \right\}.$$

Hence logistic classification is a linear classifier.

There are several ways to motivate logistic regression: for simplicity let $\mathcal{Y} = \{1, 2\}$.

1. Consider a linear regression

$$\Pr(y=1|x) = \beta_0 + x^\top \beta.$$

It violates that the constraint

$$\Pr(y=1|x) \in [0, 1].$$

A simple remedy for this problem is to set

$$\Pr(y=1|x) = F(\beta_0 + x^\top \beta),$$

where F is a continuous and strictly increasing function with $F(-\infty) = 0$ and $F(\infty) = 1$. Some examples for F are:

- Gaussian: Probit model
- Gompertz: $F(x) = \exp(-\exp(x))$, popularly used in Insurance
- Logistic: $F(x) = \exp(x)/(1+\exp(x))$.

2. Consider the decision boundary

$$\{x : \Pr(Y=1|X=x) = 0.5\}.$$

This is equivalent to

$$\{x : \log(\Pr(Y=1|X=x)/\Pr(Y=2|X=x)) = 0\}.$$

Suppose that the log-odds is linear. That is,

$$\log(\Pr(Y=1|X=x)/\Pr(Y=2|X=x)) = \beta_0 + x^\top \beta,$$

and this implies that

$$\Pr(Y=1|X=x) = \frac{\exp(\beta_0 + x^\top \beta)}{1 + \exp(\beta_0 + x^\top \beta)}.$$

For the estimation, we use the maximum likelihood approach. The likelihood is simply the probability of the observations given as

$$L(\beta_0, \beta) = \prod_{i=1}^{n} \Pr(y = y_i | x = x_i),$$

and we estimate β by maximizing the log-likelihood. For two-class case where $\mathcal{Y} = \{1, 2\}$, the log-likelihood is simplified as

$$l(\beta_{10}, \beta_1) = \sum_{i=1}^{n} \left(I(y_i = 1)(\beta_{10} + x_i^\top \beta_1) - \log(1 + \exp(\beta_{10} + x_i^\top \beta_1)) \right).$$

One obstacle of using the logistic regression would be computation since maximizing the log-likelihood is not easy. We do it by using the Iteratively Reweighted Least Squares (IRLS) algorithm.

4 Linear Discriminant Analysis (선형판별분석)

Let $f_j(x)$ is the class conditional density of x in class $y = j$ where $y \in \mathcal{Y} = \{1, \ldots, K\}$, i.e.,

$$f_j(x) = p(x|y = j), \qquad j = 1, \ldots, K.$$

Let

$$\pi_j = \Pr(y = j), \qquad j = 1, \ldots, K,$$

be the prior probabilities, with $\sum_{j=1}^{K} \pi_j = 1$. Recall that the Bayes classifier is

$$G(x) = \arg \max_{j=1,\ldots,K} P(Y = j|X = x).$$

Then Bayes theorem gives us

$$P(Y = j|X = x) = \frac{f_j(x)\pi_j}{\sum_{l=1}^{K} f_l(x)\pi_l}.$$

And hence the Bayes classifier is given as

$$G(x) = \arg \max_{j=1,\ldots,K} P(Y = j|X = x) = \arg \max_{j=1,\ldots,K} f_j(x)\pi_j.$$

Suppose that we model each class density as multivariate Gaussian

$$f_j(x) = \frac{1}{(2\pi)^{p/2} |\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_j)^\top \Sigma_j^{-1}(x - \mu_j)\right),$$

where μ_j is the mean vector and Σ_j is the covariance matrix. Then

$$\log\left(f_j(x)\pi_j\right) = -\frac{1}{2}\log|\Sigma_j| - \frac{1}{2}(x - \mu_j)^\top \Sigma_j^{-1}(x - \mu_j) + \log \pi_j - \frac{p}{2}\log(2\pi).$$

Hence by letting

$$\delta_j(x) \coloneqq -\frac{1}{2}\log|\Sigma_j| - \frac{1}{2}(x - \mu_j)^\top \Sigma_j^{-1}(x - \mu_j) + \log \pi_j,$$

then the Bayes classifier is given as

$$G(x) = j \text{ such that } \delta_j(x) > \delta_l(x) \text{ for all } l \neq j.$$

We call the functions $\delta_j(x)$ the discriminant functions.

For Linear Discriminant Analysis (LDA), we assume that

$$\Sigma_k = \Sigma \text{ for all } k.$$

In this case, we can see that

$$\begin{aligned}\log \frac{P(Y = j|X = x)}{P(Y = l|X = x)} &= \log \frac{f_j(x)}{f_l(x)} + \log \frac{\pi_j}{\pi_l} \\ &= \log \frac{\pi_j}{\pi_l} - \frac{1}{2}(\mu_j + \mu_l)^\top \Sigma(\mu_j - \mu_l) + x^\top \Sigma(\mu_j - \mu_l).\end{aligned}$$

Hence the Bayes classifier is given as

$$G(x) = j \text{ such that } \delta_j(x) > \delta_l(x) \text{ for all } l \neq j,$$

where

$$\delta_j(x) \coloneqq x^\top \Sigma^{-1}\mu_j - \frac{1}{2}\mu_j^\top \Sigma \mu_j + \log \pi_j.$$

That is, the Bayes classifier is a linear classifier. The functions δ_j are called the linear discriminant functions. If we don't assume $\Sigma_k = \Sigma$ for all k, then the decision boundary of the Bayes classifier is a quadratic function. This is called Quadratic Discriminant Analysis (QDA).

We can easily estimate μ_j and Σ_j by

- $\hat{\pi}_j = n_j/n$, where $n_j = \sum_{i=1}^n I(y_i = j)$.
- $\hat{\mu}_j = \sum_{i=1}^n x_i I(y_i = j)/n_j$.
- $\hat{\Sigma}_j = \sum_{i=1}^n (x_i - \hat{\mu}_j)(x_i - \hat{\mu}_j)^\top I(y_i = j)/(n_j - 1)$.

And we estimate Σ by the pooled variance-covariance matrix

$$\hat{\Sigma} = \frac{1}{n-K} \sum_{j=1}^K (n_j - 1)\hat{\Sigma}_j.$$

5 LDA or Logistic Classification

- Logistic classification and LDA both have linear decision boundaries.
- Logistic classification only needs the specification of the conditional distribution $\Pr(Y = j|X = x)$, that is, $\Pr(X = x)$ is completely undetermined. On the other hand, the LDA needs the specification of the joint distribution $\Pr(Y, X)$. In fact, in LDA, the marginal distribution of x is a mixture of Gaussians

  $$\Pr(x) = \sum_{j=1}^K \pi_j N(\mu_j, \Sigma).$$

  Hence, LDA needs more assumptions and hence less applicability than the logistic regression.
- Categorical input variables are allowable for the logistic regression (using dummy variables) while LDA has troubles with such inputs.
- However, LDA is a useful tool when some of the output are missing (semi-supervised learning).
- LDA is useful when Gaussian assumptions are reasonable.
- LDA works better for multi-class problems ($K > 2$).
- In practice, for a two-class problem, logistic classification and LDA are often very similar.

5.1 Multi-class problems with regression approach

There is a serious problem with the regression approach when the number of classes $K \geq 3$, especially prevalent when K is large. Because of the rigid nature of the regression model, classes can be masked by others. Figure 1 illustrates an extreme situation when $K = 3$. The three classes are perfectly separated by linear decision boundaries, yet linear regression misses the middle class completely.

In Figure 2 we have projected the data onto the line joining the three centroids (there is no information in the orthogonal direction in this case), and we have included and coded the three response variables Y_1, Y_2, and Y_3. The three regression lines (left panel) are included, and we see that the line corresponding to the middle class is horizontal and its fitted values are never dominant! Thus, observations from class 2 are classified either as class 1 or class 3. The right panel uses quadratic regression rather than linear regression. For this simple example a quadratic rather than linear fit (for the middle class at least) would solve the problem, but in general, if $K \geq 3$ classes are lined up, polynomial terms up to degree $K - 1$ might be needed to resolve them.

Note: masking problem is severe in ordinary regression (that is, regress Y on X), but it is also present in logistic regression as well. For tackling masking problem, LDA is better than logistic classification.

References

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning.* Springer Series in Statistics. Springer, New York, second edition, 2009. Data mining, inference, and prediction.

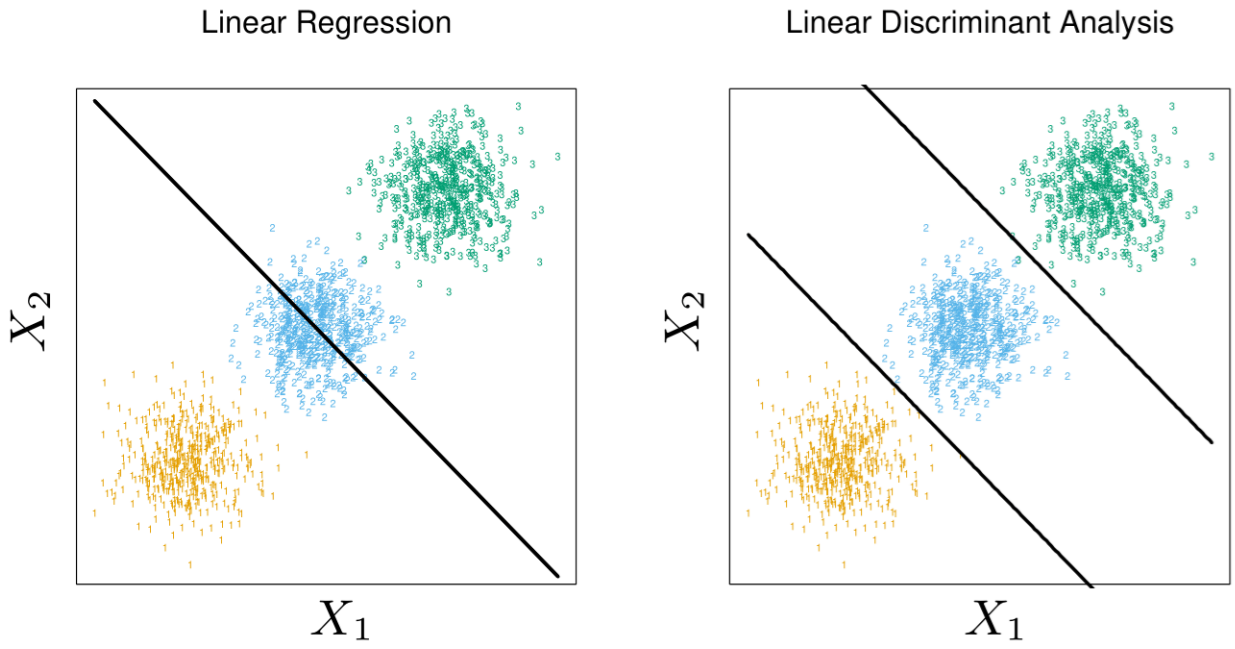

Figure 1: The data come from three classes in $\mathbb{R}^2$ and are easily separated by linear decision boundaries. The right plot shows the boundaries found by linear discriminant analysis. The left plot shows the boundaries found by linear regression of the indicator response variables. The middle class is completely masked (never dominates). Figure 4.2 from [1].

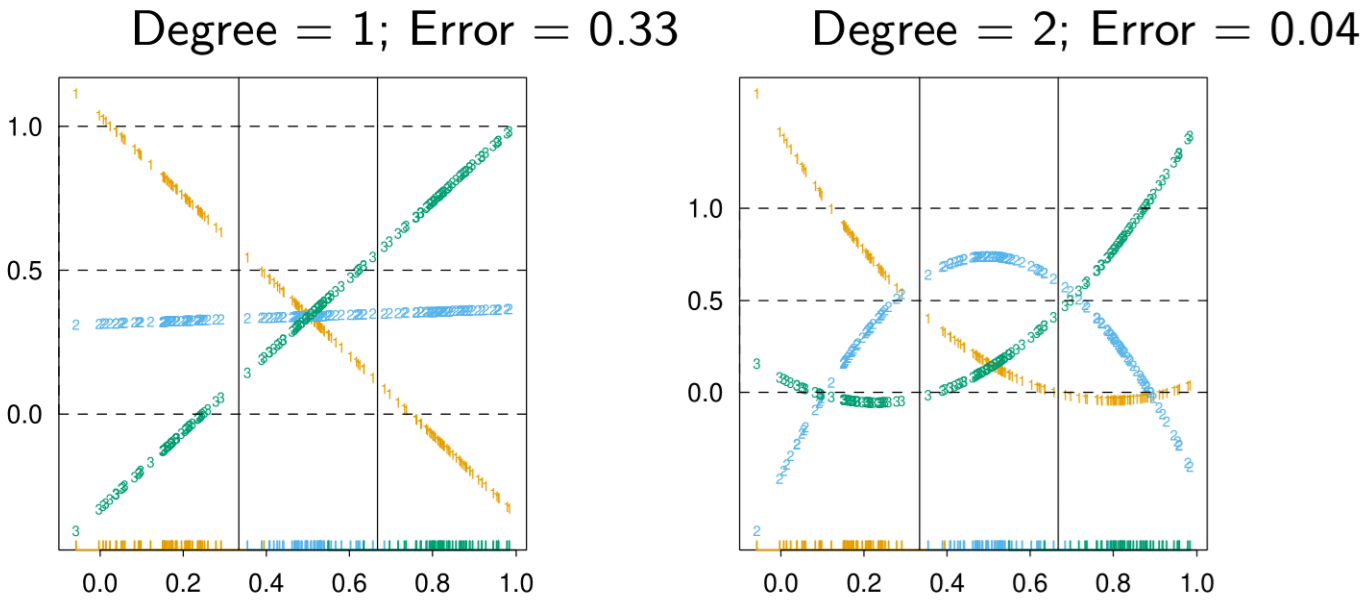

Figure 2: The effects of masking on linear regression in IR for a three-class problem. The rug plot at the base indicates the positions and class membership of each observation. The three curves in each panel are the fitted regressions to the three-class indicator variables; for example, for the blue class, yblue is 1 for the blue observations, and 0 for the green and orange. The fits are linear and quadratic polynomials. Above each plot is the training error rate. The Bayes error rate is 0.025 for this problem, as is the LDA error rate. Figure 4.3 from [1].