From Predictive to Generative Modeling

Lecturer: Qiang Sun — Lecture 03
February, 2026

Contents

1 High Dimensional Data

We first give an intuitive definition of high dimensional data.

Definition 1.1 (High Dimensional Data)

Data with "a lot of" features. More specifically, when the dimensionality d is comparable to (or even much larger than) the sample size n, it is called high dimensional data.

Let us use a concrete example to illustrate the effect of high dimensional data. Recall that

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \in \mathbb{R}^n, \quad \mathbf{X} = \begin{bmatrix} X_{11} & \dots & X_{1d} \\ \vdots & & \vdots \\ X_{n1} & \dots & X_{nd} \end{bmatrix} \in \mathbb{R}^{n \times d}.$$

We may ask the following question.

Question 1. what will happen to the OLS estimator $\widehat{\beta}^{\mathrm{OLS}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$ when $d > n$?

When $d > n$, $\mathbf{X}^\top \mathbf{X}$ is not invertible, which is due to the reason that the model or the hypothesis is too large because the dimensionality d is too large. This results in a large parameter space. Now the question is what we can do to deal with this problem. The answer is to use regularization to reduce the model/hypothesis complexity. To this end, we generally have two ways as follows.

1. Two-step procedures: reduce the dimensionality first and then regress Y on the reduced features. For example, we could apply principal component analysis to X and then regress Y on the first several principal components. Its potential drawback is that the predictors are possibly lost in the dimension reduction step.

2. Single-step regularized methods: A famous example is the Ridge estimator.

2 The Ridge Estimator

The ridge regression adds an extra ℓ_2 ball constraint to the parameter set to reduce the model complexity. The ridge regression estimator is defined as

$$\widehat{\beta}^t = \operatorname*{argmin}_{\|\beta\|_2^2 \le t} \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 \right\}, \tag{Primal Formulation}$$

where we imposed an ℓ_2-regularization $\|\beta\|_2^2 \le t$ on the parameter space. It is equivalent to the following penalized estimator

$$\widehat{\beta}^{\lambda} = \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\}. \tag{Dual Formulation}$$

The equivalence is characterized in the following lemma.

Lemma 2.1

For any $\lambda > 0$, there exists a one-to-one mapping $t = t(\lambda)$ such that

$$\widehat{\beta}^{\lambda} = \widehat{\beta}^{t}.$$

Remark 2.1

The ridge regression estimator in the dual formulation has a closed form representation as

$$\widehat{\beta}^{\lambda} = (\mathbf{X}^\top \mathbf{X} + \lambda I)^{-1} \mathbf{X}^\top \mathbf{Y}.$$

Note that the matrix $\mathbf{X}^\top \mathbf{X} + \lambda I$ is always invertible.

Remark 2.2

When t is big enough such that $t > \|\widehat{\beta}^{\mathrm{OLS}}\|_2^2$, this corresponds to $\lambda = 0$. If $t = 0$, then this corresponds to $\lambda = \infty$.

Example 2.1

Suppose $\mathbf{X}^\top \mathbf{X} = I$. Then we have

$$\widehat{\beta}^{t} = \widehat{\beta}^{\lambda} = (\mathbf{X}^\top \mathbf{X} + \lambda I)^{-1} \mathbf{X}^\top \mathbf{Y} = (I + \lambda I)^{-1} \mathbf{X}^\top \mathbf{Y} = \frac{1}{1+\lambda} \mathbf{X}^\top \mathbf{Y} = \frac{1}{1+\lambda} \widehat{\beta}^{\mathrm{OLS}}.$$

Definition 2.2 (Contour)

A contour line of a function is a curve along which the function has a constant value.

Figure 2 shows the geometric interpretation of the ridge regression estimator with $d = 2$, where the blue curves are the contour lines of the least square objection function $F(\beta) := \|\mathbf{Y} - \mathbf{X}^\top \beta\|_2^2$. β^{λ} has a closed form solution and it is convex, but we should never use generic solvers.

Then we may have a question about the model selection. According to Lemma 2.1, each λ index a model in $\mathcal{P}_{\lambda}\{Y = \beta^\top X + \epsilon : \|\beta\|_2^2 \le t(\lambda),\ \epsilon \sim \mathcal{N}(0, \sigma^2),\ \mathbb{E}[\epsilon | X] = 0\}$, so the model selection is equivalent to the selection of the tuning parameter λ.

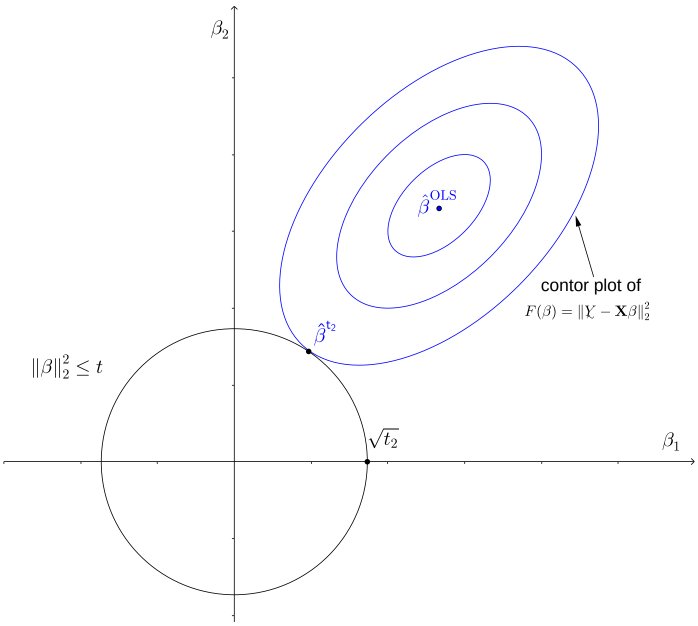
β_2
$\widehat{\beta}^{\mathrm{OLS}}$
contor plot of
$F(\beta) = \|\underset{\sim}{Y} - \mathbf{X}\beta\|_2^2$
$\widehat{\beta}^{t_2}$
$\|\beta\|_2^2 \leq t$
$\sqrt{t_2}$
β_1

Figure 1: The geometric interpretation for the ridge regression estimator of $\beta = (\beta_1, \beta_2)^\top$.

3 Model selection

A basic method for model selection is data splitting. We split the data $\mathcal{D} = \{X_1,\ X_2,\ \ldots,\ X_n\}$ into two subsets $\mathcal{D}_1$ and $\mathcal{D}_2$ with sizes n_1 and n_2 such that $\mathcal{D}_1 \bigcup \mathcal{D}_2 = \mathcal{D}$ and $n_1 + n_2 = n$. Consider a pool of tuning parameter

$$\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_K\}.$$

Let $\widehat{\beta}^{\lambda_1}, \ldots, \widehat{\beta}^{\lambda_K}$ be the ridge regression estimators on the subset $\mathcal{D}_1$. We define the data splitting score corresponding to λ_k as

$$\mathcal{DS}(k) = \frac{1}{n_2} \sum_{i \in \mathcal{D}_2} (Y_i - X_i^\top \widehat{\beta}^{\lambda_k})^2, \quad 1 \leq k \leq K.$$

We then pick the λ_k/model with the smallest $\mathcal{DS}$ score. The basic theory behind this method is: conditioning on $\mathcal{D}_1$, it is easy to see that $\mathcal{DS}(k)$ is an unibiased estimator of the risk $R(\widehat{\beta}^{\lambda_k})$. We summarize the advantages and disadvantages of data splitting as follows.

1. Data splitting has good generalization performance and it is theoretically and computationally simple.

2. Data splitting may cause a "waste" of training data, because the validation subset $\mathcal{D}_2$ is not used for training at all.

Then we introduce the cross validation (CV) as a solution to the drawback of the data splitting.

Definition 3.1 (J-Fold Cross Validation)

We split the data $\mathcal{D}$ into J equal sized parts/subsets $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_J$. This forms J binary splits as follows.

$$\begin{array}{cccc} \mathcal{DS}_1 & \mathcal{D}_1 & \text{versus} & \mathcal{D}\backslash\mathcal{D}_1 \\ \mathcal{DS}_2 & \mathcal{D}_2 & \text{versus} & \mathcal{D}\backslash\mathcal{D}_2 \\ \vdots & \vdots & \vdots & \vdots \\ \mathcal{DS}_J & \mathcal{D}_J & \text{versus} & \mathcal{D}\backslash\mathcal{D}_J \end{array}$$

For each $\lambda_k \in \Lambda$, we calculate the data splitting scores $\mathcal{DS}_1, \mathcal{DS}_2, \ldots, \mathcal{DS}_J$ and denote the results as $\mathcal{DS}_1(k), \mathcal{DS}_2(k), \ldots, \mathcal{DS}_J(k)$. We define the cross validation score as

$$\mathrm{CV}(k) = \frac{1}{J} \sum_{j=1}^{J} \mathcal{DS}_j(k).$$

We then pick the λ_j/model with the smallest CV score.

Remark 3.1

After picking λ_j, we could use this λ_j to fit the entire data set $\mathcal{D}$.

4 Bridge Regression Estimator

The ridge regression reduces the model complexity by imposing an ℓ_2 norm constraint. Bridge estimators use a general ℓ_p constraint to reduce the model complexity.

Let $x = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$, ℓ_p-norm ($p \geq 1$) of x is defined as

$$\|x\|_p = (|x_1|^p + \ldots + |x_n|^p)^{1/p}.$$

For example, when $p = 2$, we have the ℓ_2 norm $\|x\|_2 = (x_1^2 + \ldots + x_n^2)^{1/2}$. We have two important observations.

1. For $1 \leq p < \infty$, $\|x\|_p$ is a norm, and thus $\{x : \|x\|_p \leq t\}$ is a convex constraint.
2. For $0 < p < 1$, $\|x\|_p$ is not a norm and then $\{x : \|x\|_p \leq t\}$ is not a convex constraint.

Figure 2 shows the geometric interpretation of the ℓ_p constraint for different p's.

Replacing the ℓ_2 regularization by a general ℓ_p regularization, we define the bridge estimator as follows.

Definition 4.1 (The Bridge Estimator Family)

For any $0 < p < \infty$ and $\lambda > 0$, the bridge estimator is defined as

$$\widehat{\beta}^{\text{bridge}} = \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_p^p \right\}.$$

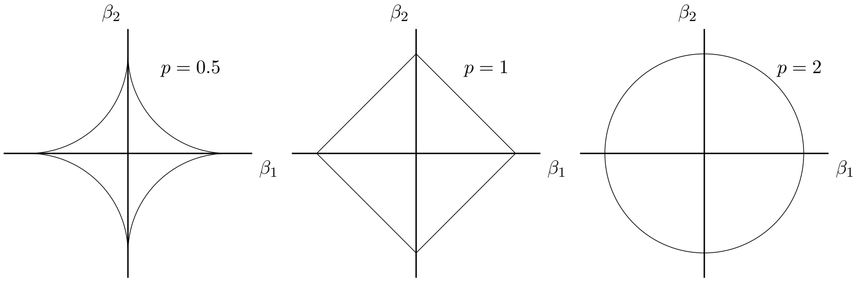

Figure 2: ℓ_p-ball constraint on $\beta = (\beta_1, \beta_2)^\top$ with $p = 0.5$, 1 and 2 respectively.

The ridge estimator includes many estimator as special cases. For example, when taking $p = 2$, then the bridge estimator reduces to the ridge estimator. In the bridge estimator family, $p = 1$ is the most important case which leads to the well-known the *least absolute shrinkage and selection operator* (Lasso) estimator.

5 The Lasso Estimator

Definition 5.1 (Lasso Estimator)

For any $t > 0$, the Lasso estimator is defined as

$$\widehat{\beta}^t = \underset{\|\beta\|_1 \le t}{\operatorname{argmin}} \{\|\mathbf{Y} - \mathbf{X}\beta\|_2^2\}. \qquad \text{(Primal Formulation)}$$

The equivalent penalized estimator formulation is

$$\widehat{\beta}^\lambda = \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \{\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1\}, . \qquad \text{(Dual Formulation)}$$

where $\lambda > 0$ characterizes the amount of shrinkage.

Figure 3 shows the geometric interpretation of the Lasso estimator with $d = 2$, where the blue curves are the contour lines of the least square objection function $F(\beta) := \|\mathbf{Y} - \mathbf{X}^\top\beta\|_2^2$. In this example, we have $\widehat{\beta}_1^{\text{Lasso}} = 0$ and $\widehat{\beta}_2^{\text{Lasso}} \neq 0$. A more general fact is that the Lasso estimator has a *sparsity* property. Here we use “sparsity” to refer to that many elements of β are 0. Thus, the Lasso estimator could result in variable selections. We use an example to illustrate this.

Example 5.1

Suppose we have a regression function $f(X) = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_d X_d$ with $\beta_2 = 0$. Then β_2 does not contribute to the prediction at all. The Lasso estimator with a proper parameter $\lambda > 0$ could lead to variable selection: $\widehat{\beta}_2^\lambda = 0$.

For the selection of the tuning parameter λ in a Lasso estimator, we can use data splitting or cross validation as the same in ridge estimators. We briefly repeat the procedure of data splitting.

Suppose we have a dataset $\mathcal{D} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. We split the dataset $\mathcal{D}$ into

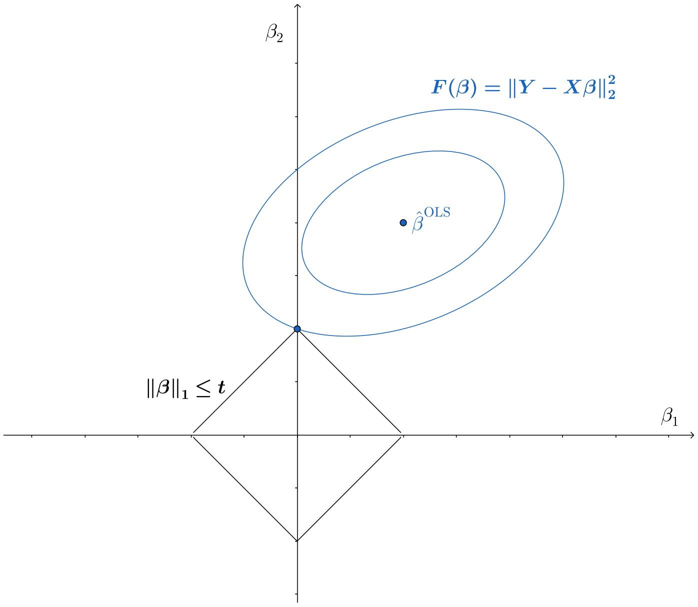

Figure 3: The geometric interpretation for the ridge regression estimator of $\beta = (\beta_1, \beta_2)^\top$.

two subsets $\mathcal{D}_1$ and $\mathcal{D}_2$ with sizes n_1 and n_2 such that $\mathcal{D}_1 \bigcup \mathcal{D}_2 = \mathcal{D}$ and $n_1 + n_2 = n$, which implies $\mathcal{D}_1 \bigcap \mathcal{D}_2 = \emptyset$. Consider a pool of tuning parameter

$$\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_K\}.$$

We first learn $\widehat{\beta}^{\lambda_1}, \ldots, \widehat{\beta}^{\lambda_K}$ from $\mathcal{D}_1$. Then we evaluate $\widehat{\beta}^{\lambda_1}, \ldots, \widehat{\beta}^{\lambda_K}$ on $\mathcal{D}_2$ by comparing their data splitting scores

$$\mathcal{DS}(k) = \frac{1}{n_2} \sum_{i \in \mathcal{D}_2} (Y_i - X_i^\top \widehat{\beta}^{\lambda_k})^2, \quad 1 \leq k \leq K.$$

We then choose the λ_k with the minimal $\mathcal{DS}$ score.

Remark 5.1

A larger λ is more likely to produce a sparse solution/estimator.

Remark 5.2

In bridge estimator family, $0 < p \leq 1$ corresponds to sparsity and $1 \leq p < \infty$ corresponds to convexity. This is the reason why the Lasso estimator ($p = 1$) is the most important case in the family.

6 The Elastic Net

We first recall some facts for ridge estimators and Lasso estimators.

1. A ridge estimator is not sparse but it corresponds to a strongly convex optimization problem and handles collinearity. Here the collinearity refers to a phenomenon that two or more predictor variables are highly correlated.

2. A Lasso estimator is sparse and convex, but it could not handle collinearity well.

By combining the advantages of ridge estimators and Lasso estimators, we introduce the elastic net as follows.

Definition 6.1 (Elastic Net)

For any $\lambda > 0$ and $0 \leq \alpha \leq 1$, the elastic net is defined as

$$\widehat{\beta}^{\text{Elastic}} = \underset{\beta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \left(\alpha \|\beta\|_1 + (1-\alpha)\|\beta\|_2^2 \right) \right\},$$

which reduces to the Lasso estimator if $\alpha = 1$ or the ridge estimator if $\alpha = 0$.

Remark 6.1

For the selection of the tuning parameter λ in an elastic net, we can also use data splitting or cross validation as the same in lasso estimators and ridge estimators.

7 From linear to nonlinear regression: Basis expansion

Linear regression might be considered as very restrictive given the linear hypothesis. But contrary to popular belief, linear regression models are very flexible. The process of feature engineering allows us to improve the model complexity, and thus increase the fitting power. We show some examples on constructing nonlinear features as follows.

1. Input can be transformations of original features. Examples are

$$X_1' = \log(X_1), \quad X_1' = \sqrt{X_1}, \quad X_1' = X_1^2. \qquad \text{(Handcrafted features)}$$

2. Inputs can be interaction terms. Based on the original features $X_1, \ldots, X_d$, we have order-2 interactions

$$X_1X_2,\ X_1X_3, \ldots, X_dX_{d-1}.$$

3. Input can have basis expansions. Instead of using

$$f(X) = \sum_{j=1}^{d} \beta_j X_j,$$

we consider

$$f(X) = \sum_{j=1}^{d} \beta_j h_j(X)$$

where h_j's are some suitably chosen functions, such as wavelet functions in image analysis and polynomial basis function $\{h(x) = x^a : a = 0, 1, 2, 3, \ldots\}$ in nonparametric statistics.

4. Input can be indicator functions of qualitative inputs

$$I(X_j \in A) = \begin{cases} 1 & X_j \in A, \\ 0 & X_j \notin A, \end{cases}$$

where A is a set of qualitative values.

More general than item 4 above, we can consider categorical data analysis, which is a set of analysis techniques that are used in supervised learning with categorical variables. First, let us define categorical variables.

Definition 7.1 (Categorical Variable)

A categorical variable is a variable that can only take on one of a limited values.

Example 7.1 (Dummy Coding for Creating Categorical Variables)

To denote the gender of a person, a categorical variable can be created using dummy coding. Dummy encoding proceeds by assigning a quantitive value to a qualitative value. For example, if X describes the gender, using dummy coding we can create categorical variables as

$$I(X = \text{male}) = \begin{cases} 1 & X = \text{male}, \\ 0 & X = \text{female}. \end{cases}$$

More generally, if a categorical variable has K categories. We can create $K-1$ categorical variables using the following dummy coding:

	X_1	X_2	$\dots$	X_{K-1}
City 1	0	0	$\dots$	0
City 2	1	0	$\dots$	0
City 3	0	1	$\dots$	0
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
City K	0	0	$\dots$	1

Then our $K-1$ dummy variables are

$$I(X_j = \text{City } j+1), \text{ for } 1 \leq j \leq K-1.$$

Note that the above dummy coding has no ordinal information which means we do not assume that K cities are ordered in anyway: it is a general coding.

8 Generative Modeling

We first recall the logistic regression which directly models

$$\mathbb{P}(Y = y | X = x) = \frac{1}{1 + e^{-yf(x)}}, \ y \in \{-1, +1\}.$$

Another way to model the conditional probability $\mathbb{P}(Y = y|X = x)$ is by Bayes formula

$$\begin{aligned}\mathbb{P}(Y = +1|X = x) &= \frac{\mathbb{P}(x|Y = +1)\mathbb{P}(Y = +1)}{\mathbb{P}(x)} \\ &= \frac{\mathbb{P}(x|Y = +1)\mathbb{P}(Y = +1)}{\mathbb{P}(x|Y = +1)\mathbb{P}(Y = +1) + \mathbb{P}(x|Y = -1)\mathbb{P}(Y = -1)}.\end{aligned}$$

Defining

$$\mathbb{P}(Y = +1) =: \eta, \quad \mathbb{P}(x|Y = +1) =: p_+(x), \quad \mathbb{P}(x|Y = -1) =: p_-(x),$$

we have

$$\mathbb{P}(Y = +1|X = x) = \frac{p_+(x)\eta}{p_+(x)\eta + p_-(x)(1-\eta)}$$

and

$$\mathbb{P}(Y = -1|X = x) = \frac{p_-(x)(1-\eta)}{p_+(x)\eta + p_-(x)(1-\eta)}.$$

Note that the conditional probability $\mathbb{P}(Y = +1|X = x)$ is the key quantity to get the Bayes rule $h^*(x)$, since

$$h^*(x) = \begin{cases} +1, & \text{if } \mathbb{P}(Y = +1|X = x) > \frac{1}{2}; \\ -1, & \text{otherwise.} \end{cases}$$

To this end, we only need to estimate η, $p_+(x)$ and $p_-(x)$ given i.i.d. data $(X_1, Y_1), \ldots, (X_n, Y_n)$. We consider the MLE of η, $p_+(x)$ and $p_-(x)$ as follows.

We first have the log-likelihood function

$$\begin{aligned}&\ell_n(\eta, p_+, p_-|X_1, \ldots, X_n; Y_1, \ldots, Y_n) \\ =&\sum_{i=1}^n \log \mathbb{P}(X_i, Y_i) \\ =&\sum_{i:Y_i=+1} \log \mathbb{P}(X_i, Y_i) + \sum_{i:Y_i=-1} \log \mathbb{P}(X_i, Y_i) \\ =&\sum_{i:Y_i=+1} \log \mathbb{P}(X_i|Y_i = +1) + \sum_{i:Y_i=+1} \log \eta + \sum_{i:Y_i=-1} \log \mathbb{P}(X_i|Y_i = -1) + \sum_{i:Y_i=-1} \log(1-\eta).\end{aligned}$$

To optimize over η, we denote $n_+ = \sum_{i=1}^n \mathbb{I}(Y_i = +1)$ and $n_- = n - n_+$. Then it holds that

$$\widehat{\eta}^{\text{MLE}} = \operatorname*{argmax}_{\eta} \ell_n(\eta, p_+, p_-) = \operatorname*{argmax}_{\eta} \left(\sum_{i:Y_i=+1} \log \eta + \sum_{i:Y_i=-1} \log(1-\eta) \right) = \frac{n_+}{n},$$

where we used the fact that

$$\frac{\partial \ell_n}{\partial \eta} = \frac{\partial}{\partial \eta} \left(\sum_{i:Y_i=+1} \log \eta + \sum_{i:Y_i=-1} \log(1-\eta) \right) = \frac{n_+}{\eta} - \frac{n_-}{1-\eta}.$$

9 Discriminant Analysis

9.1 Quadratic Discriminant Analysis

To optimize $p_+(x)$ and $p_-(x)$, we could consider the *Gaussian Discriminant Analysis* (or *Quadratic Discriminant Analysis*).

Definition 9.1 (Quadratic Discriminant Analysis (QDA))

Quadratic Discriminant Analysis refers to modeling $p_+(x)$ and $p_-(x)$ by

$$p_+(x) = \frac{1}{(2\pi)^{d/2}|\Sigma_+|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_+)^\top \Sigma_+^{-1}(x-\mu_+)\right\}, \tag{9.1}$$

$$p_-(x) = \frac{1}{(2\pi)^{d/2}|\Sigma_-|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_-)^\top \Sigma_+^{-1}(x-\mu_-)\right\} \tag{9.2}$$

with mean vectors $\mu_+, \mu_- \in \mathbb{R}^d$ and population covariance matrices $\Sigma_+, \Sigma_- \in \mathbb{R}^{d\times d}$.

Recall the Bayes rule that

$$h^*(x) = \begin{cases} +1 & \text{if } \mathbb{P}(Y=+1|X=x) > \mathbb{P}(Y=-1|X=x), \\ -1 & \text{otherwise.} \end{cases}$$

Under the Gaussian assumptions (9.1) and (9.2), the condition $\mathbb{P}(Y=+1|X=x) > \mathbb{P}(Y=-1|X=x)$ becomes

$$\frac{p_+(x)\eta}{p_+(x)\eta + p_-(x)(1-\eta)} > \frac{p_-(x)(1-\eta)}{p_+(x)\eta + p_-(x)(1-\eta)},$$

After some basic algebra, it is equivalent to

$$\frac{1}{2}\log\frac{|\Sigma_-|}{|\Sigma_+|} + \frac{1}{2}(x-\mu_-)^\top\Sigma_-(x-\mu_-) - \frac{1}{2}(x-\mu_+)^\top\Sigma_+(x-\mu_+) + \log\frac{\eta}{1-\eta} > 0. \tag{9.3}$$

Then the Bayes classification rule of GDA/QDA turns to

$$h^*(x) = \begin{cases} +1 & \text{if } \frac{1}{2}r_-^2(x) - \frac{1}{2}r_+(x)^2 + \frac{1}{2}\log\frac{|\Sigma_-|}{|\Sigma_+|} + \log\frac{\eta}{1-\eta} > 0, \\ -1 & \text{otherwise,} \end{cases} \tag{9.4}$$

where $r_-(x) := \sqrt{(x-\mu_-)^\top\Sigma_-^{-1}(x-\mu_-)}$ and $r_+(x) := \sqrt{(x-\mu_+)^\top\Sigma_+^{-1}(x-\mu_+)}$. Here $r_-(x)$ (resp. $r_+(x)$) is called *Mahalanobis distance* of the observation $x \in \mathbb{R}^d$ with mean μ_- (resp. μ_+) and covariance matrix Σ_- (resp. Σ_+).

In the following proposition, we show the MLE of parameters μ_-, μ_+, Σ_- and Σ_+ in the QDA model.

Proposition 9.2

Let $n_+ = \sum_{1\le i\le n}\mathbb{I}(Y_i = 1)$ and $n_- = \sum_{1\le i\le n}\mathbb{I}(Y_i = -1)$. Under the QDA model

(9.1) and (9.2), one has

$$\widehat{\mu}_+^{\mathrm{MLE}} = \frac{1}{n_+} \sum_{i:Y_i=+1} X_i, \quad \widehat{\mu}_-^{\mathrm{MLE}} = \frac{1}{n_-} \sum_{i:Y_i=-1} X_i,$$

$$\widehat{\Sigma}_+^{\mathrm{MLE}} = \frac{1}{n_+} \sum_{i:Y_i=+1} (X_i - \widehat{\mu}_+^{\mathrm{MLE}})^\top (X_i - \widehat{\mu}_+^{\mathrm{MLE}}),$$

$$\widehat{\Sigma}_-^{\mathrm{MLE}} = \frac{1}{n_-} \sum_{i:Y_i=-1} (X_i - \widehat{\mu}_-^{\mathrm{MLE}})^\top (X_i - \widehat{\mu}_-^{\mathrm{MLE}}).$$

Remark 9.1

Note that the decision boundary in this case is $\{x : \frac{1}{2}r_-^2(x) - \frac{1}{2}r_+(x)^2 + \frac{1}{2}\log\frac{|\Sigma_-|}{|\Sigma_+|} + \log\frac{\eta}{1-\eta} = 0\}$ which includes two quadratic functions $\frac{1}{2}r_-^2(x)$ and $\frac{1}{2}r_+^2(x)$. This is the reason why we call it Quadratic Discriminant Analysis (QDA). When we take $\Sigma_- = \Sigma_+$, the QDA reduces to the special case Linear Discriminant Analysis (LDA).

9.2 Linear Discriminant Analysis

Definition 9.3 (Linear Discriminant Analysis (LDA))

Linear Discriminant Analysis refers to the special case of QDA with the regularization $\Sigma_+ = \Sigma_- = \Sigma$.

Figure 4 shows an example of binary classification with two linearly separable sets of observations so that the LDA could work perfectly.

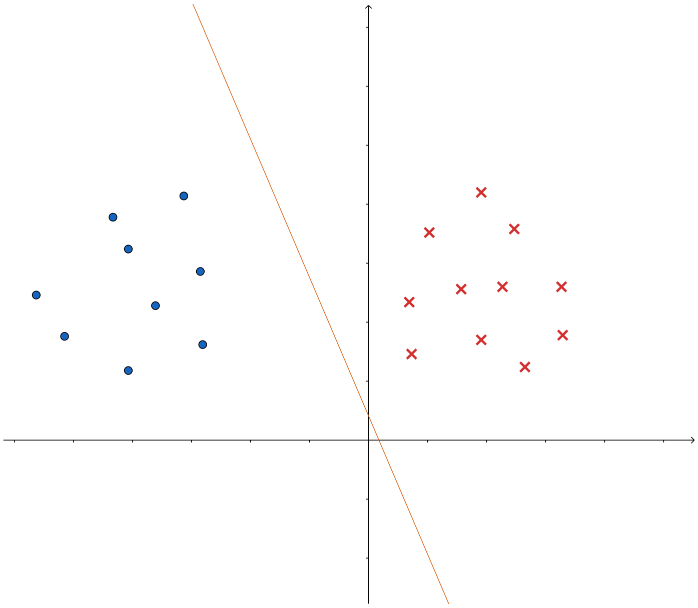

Figure 4: An example of binary classification where the LDA works perfectly.

A natural question arises: what is the Bayes classification rule for LDA? According to the assumption that $\Sigma_+ = \Sigma_- = \Sigma$ made for LDA, the condition (9.3) can be simplified as

$$\frac{1}{2}(x-\mu_-)^\top \Sigma (x-\mu_-) - \frac{1}{2}(x-\mu_+)^\top \Sigma (x-\mu_+) + \log\frac{\eta}{1-\eta} > 0,$$

or equivalently,

$$(\mu_+ - \mu_-)^\top \Sigma^{-1} x + \frac{1}{2}\mu_-^\top \Sigma^{-1}\mu_- - \frac{1}{2}\mu_+^\top \Sigma^{-1}\mu_+ + \log\frac{\eta}{1-\eta} > 0. \tag{9.5}$$

It worth noting that (9.5) implies a linear classification boundary of LDA, as the orange line shown in the illustrative example in Figure 4. More specifically, this linear classification boundary is $\beta^\top x + \beta_0 > 0$ with $\beta := (\mu_+ - \mu_-)^\top \Sigma^{-1}$ and $\beta_0 := \frac{1}{2}\mu_-^\top \Sigma^{-1}\mu_- - \frac{1}{2}\mu_+^\top \Sigma^{-1}\mu_+ + \log\frac{\eta}{1-\eta}$.

Remark 9.2

Recall that, in a linear logistic regression (LLR), we model the odd ratio $\frac{\mathbb{P}(Y=+1|X=x)}{\mathbb{P}(Y=-1|X=x)}$ by

$$\log\frac{\mathbb{P}(Y=+1|X=x)}{\mathbb{P}(Y=-1|X=x)} = f(x),$$

where $f(x)$ is restricted to the linear function class. We may note that LDA also leads to a linear model as

$$\log\frac{\mathbb{P}(Y=+1|X=x)}{\mathbb{P}(Y=-1|X=x)} = \beta^\top x + \beta_0.$$

The LDA is very similar to LLR in this sense and can be viewed as a special case of the logistic regression (LR) (or, a more regularized version of the LR), although LDA does not have the same model space with LLR. Note here that we should always compare the joint distributions but not “marginal” when comparing model space.

We could derive the MLE of μ_-, μ_+ and Σ similar to Proposition 9.2.

Proposition 9.4

For the LDA model, one has

$$\widehat{\mu}_+^{\mathrm{MLE}} = \frac{1}{n_+}\sum_{i:Y_i=+1} X_i, \quad \widehat{\mu}_-^{\mathrm{MLE}} = \frac{1}{n_-}\sum_{i:Y_i=-1} X_i, \quad \widehat{\Sigma}^{\mathrm{MLE}} = \frac{n_+\widehat{\Sigma}_+ + n_-\widehat{\Sigma}_-}{n_+ + n_-},$$

where

$$\widehat{\Sigma}_+ = \frac{1}{n_+}\sum_{i:Y_i=+1}(X_i - \widehat{\mu}_+^{\mathrm{MLE}})^\top (X_i - \widehat{\mu}_+^{\mathrm{MLE}}),$$
$$\widehat{\Sigma}_- = \frac{1}{n_-}\sum_{i:Y_i=-1}(X_i - \widehat{\mu}_-^{\mathrm{MLE}})^\top (X_i - \widehat{\mu}_-^{\mathrm{MLE}}).$$

10 Generative Modeling

We first point out the difference between generative and discriminative modeling. By Bayes formula, one has

$$p(y, x) = p(y|x)p(x).$$

The discriminative modeling will ignore the pdf $p(x)$ and only model the part of conditional probability $p(y|x)$ which is directly relevant to the classification problem, while the generative modeling will focus on both $p(x)$ and $p(y|x)$ since they as a whole affect the generation of new data.

Remark 10.1

If the LDA model is correct, the LDA method is expected to be more "efficient".

Next we consider the generative modeling in high dimensional cases. The naive Bayes modeling is a useful approach to reduce the model complexity (or, the number of parameters).

Definition 10.1 (Naive Bayes)

The *naive Bayes modeling* refers to the generative modeling with an additional "class conditional independence" regularization, i.e.,

$$\mathbb{P}(x|Y = +1) = \prod_{j=1}^{d} \mathbb{P}(x_j|Y = +1) \text{ and } \mathbb{P}(x|Y = -1) = \prod_{j=1}^{d} \mathbb{P}(x_j|Y = -1),$$

which is also called naive Bayes regularization and where $x = (x_1, \cdots, x_d)^\top \in \mathbb{R}^d$.

Under the naive Bayes regularization, we have

$$\begin{aligned}
\log \frac{\mathbb{P}(Y = +1|X = x)}{\mathbb{P}(Y = -1|X = x)} &= \log \frac{\mathbb{P}(x|Y = +1)\mathbb{P}(Y = +1)}{\mathbb{P}(x|Y = -1)\mathbb{P}(Y = -1)} \\
&= \log \frac{\mathbb{P}(x|Y = +1)}{\mathbb{P}(x|Y = -1)} + \log \frac{\eta}{1-\eta} \\
&= \log \prod_{j=1}^{d} \frac{\mathbb{P}(x_j|Y = +1)}{\mathbb{P}(x_j|Y = -1)} + \log \frac{\eta}{1-\eta} \\
&= \sum_{j=1}^{d} \log \frac{\mathbb{P}(x_j|Y = +1)}{\mathbb{P}(x_j|Y = -1)} + \log \frac{\eta}{1-\eta} \\
&=: \sum_{j=1}^{d} f_j(x_j) + \log \frac{\eta}{1-\eta}.
\end{aligned}$$

Note that $f_j(x_j)$ is a univariate function and all we need is to model the conditional probabilities $\mathbb{P}(x_j|Y = +1)$ and $\mathbb{P}(x_j|Y = -1)$ for $1 \le j \le d$. We give an example of Diagonal LDA (DLDA) for this purpose.

Example 10.1 (DLDA)

Consider a LDA model with the naive Bayes regularization

$$\Sigma = \begin{pmatrix} \sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_d^2 \end{pmatrix} \in \mathbb{R}^{d\times d}.$$

We have

$$X_j|Y=+1 \sim \mathcal{N}(\mu_{+j}, \sigma_j^2), \quad X_j|Y=-1 \sim \mathcal{N}(\mu_{-j}, \sigma_j^2), \quad \text{for } 1 \le i \le d,$$

where μ_{+j} and μ_{-j} denote the $j-$th component of μ_+ and μ_-, respectively. Then it follows that

$$\mathbb{P}(x|Y=+1) = \prod_{j=1}^{d} \mathbb{P}(x_j|Y=+1) = \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_j-\mu_j)^2}{2\sigma_j^2}\right\}.$$

Based on this, we can get the MLE of μ_+, μ_- and Σ as

$$\widehat{\mu}_+^{\text{MLE}} = \frac{1}{n_+}\sum_{i:Y_i=+1} X_i, \quad \widehat{\mu}_-^{\text{MLE}} = \frac{1}{n_-}\sum_{i:Y_i=-1} X_i, \quad \widehat{\Sigma}^{\text{MLE}} = \text{diag}(\widehat{\sigma}_1^2, \cdots, \widehat{\sigma}_d^2),$$

where

$$\widehat{\sigma}_j^2 = \frac{n_+\widehat{S}_{+j}^2 + n_-\widehat{S}_{-j}^2}{n_+ + n_-}, \; \widehat{S}_{+j}^2 = \frac{1}{n_+}\sum_{i:Y_i=+1}(X_{ij}-\widehat{\mu}_{+j})^2, \; \widehat{S}_{-j}^2 = \frac{1}{n_-}\sum_{i:Y_i=-1}(X_{ij}-\widehat{\mu}_{-j})^2.$$

Example 10.2

If X_j is a categorical variable, we can use a discrete model (e.g. Bernoulli distribution etc.).

Example 10.3 (Number of Free Parameters)

(a). For a full QDA model with parameters $(\Sigma_+, \Sigma_-, \mu_+, \mu_-, \eta)$, the total number of free parameters is $d(d+1) + 2d + 1$. Here we have $\frac{d(d+1)}{2}$ parameters in Σ_+ (or Σ_-) since $\Sigma_+ \in \mathbb{R}^{d\times d}$ is a symmetric matrix.
(b). For a full LDA model with parameters $(\Sigma, \mu_+, \mu_-, \eta)$, the total number of free parameters is $d(d+1)/2 + 2d + 1$.
(c). For a DLDA model with parameters $(\sigma_1^2, \cdots, \sigma_d^2, \mu_+, \mu_-, \eta)$, the total number of free parameters is $3d + 1$.