LINEAR AND QUADRATIC DISCRIMINANT ANALYSIS

9/3/21

We go back to classifiers (categorical responses). We have seen in detail two methods: K-nn and logistic regression. These methods focus on estimating $P(Y=j \mid X=x)$ directly.

We now consider approaches that estimate $P(Y=j \mid X=x)$ "indirectly" by modelling $P(X=x \mid Y=j)$ and then deducing $P(Y=j \mid X=x)$ from that via Bayes theorem. In particular, we will consider:

1. Linear Discriminant Analysis (LDA)
2. Quadratic Discriminant Analysis (QDA)
3. Naive Bayes Classifier

BAYES THEOREM FOR CLASS PROBABILITIES

Idea: Model the distribution of X in each class, ie $P(X=x|Y=j)$ for X discrete and $f(x|j)$ for X continuous, then use Bayes theorem to flip things around:

$$P(Y=j|X=x) = \frac{P(X=x|Y=j)\,P(Y=j)}{P(X=x)}$$

$$= \frac{P(X=x|Y=j)\,P(Y=j)}{\sum_{i=1}^{K} P(X=x|Y=i)\,P(Y=i)}$$

(K classes form a partition)

More generally (X discrete or continuous):

$$P(Y=j|X=x) = \frac{f_j(x)\,\pi_j}{\sum_{i=1}^{K} f_i(x)\,\pi_i}$$

POSTERIOR probability of a random observation belonging to class i once X is observed

where

$\pi_i = P(Y=i)$ **PRIOR** probability that a randomly chosen observation comes from class i

$$f_i(x) = \begin{cases} P(X=x|Y=i) & \text{if } X \text{ discrete} \\ f_{X|Y}(x|i) & \text{if } X \text{ continuous} \end{cases}$$

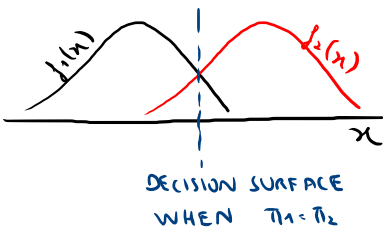
DENSITY of X for an observation that comes from class i

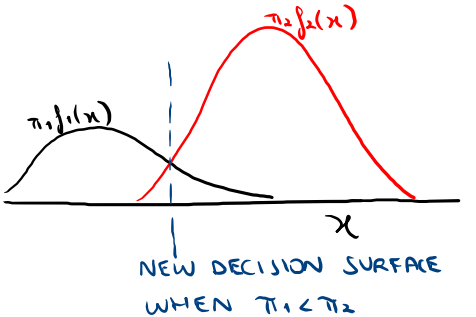
BAYES THEOREM FOR CLASSIFICATION

Using the Bayes classifier, we assign x to the class with highest posterior probability $P(Y=j \mid X=x)$.

Using Bayes theorem, and noting that the denominator is constant for all j, results in:

Assign x to class j for which $f_j(x)\pi_j$ is highest





In practice, π_j is unknown but simple to estimate, eg

① Equal for all classes: $\hat{\pi}_j = \frac{1}{K}$

② Corresponding to data frequencies: $\hat{\pi}_j = \frac{n_j}{n}$

So the focus of the modelling is to provide an estimate of $f_j(x)$ which then combined with $\hat{\pi}_j$ would give an estimate of $P(Y=j \mid x)$

LINEAR DISCRIMINANT ANALYSIS

Estimating $f_j(x)$ needs some assumptions, one of which underlies LDA.

But why do we need LDA in the first place?

1. When the classes are well separated, logistic regression estimates are (surprisingly) unstable → not LDA
2. LDA more natural when we have more than 2 classes
3. LDA better/more stable if underlying assumptions are satisfied

What are then these assumptions?

LDA FOR p = 1

Let us consider the simple case of $p = 1$ (one predictor).

LDA assumes that the density $f_j(x)$ is Gaussian for each class j.

So in the one-dimensional setting:

$$\text{①} \quad f_j(x) = \frac{1}{\sqrt{2\pi}\,\sigma_j} e^{-\frac{1}{2}\left(\frac{x-\mu_j}{\sigma_j}\right)^2}, \quad x \in \mathbb{R}$$

where μ_j is the mean and σ_j^2 the variance (given class j).

LDA makes a second assumption:

$$\text{②} \quad \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_K^2 = \sigma^2 \quad \text{(homoskedasticity)}$$

① and ② lead to

$$f_j(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_j}{\sigma}\right)^2}, \quad x \in \mathbb{R}, \quad j = 1, \ldots, K$$

In short, $\boxed{X \mid Y = j \sim N(\mu_j, \sigma^2)}$.

LDA FOR CLASSIFICATION

Using Bayes theorem with this $f_j(x)$ leads to

$$P(Y=j \mid X=x) = \frac{\cancel{\frac{1}{\sqrt{2\pi}\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu_j}{\sigma}\right)^2} \pi_j}{\sum_{i=1}^{K} \cancel{\frac{1}{\sqrt{2\pi}\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma}\right)^2} \pi_i}$$

and to assigning x to the class with the largest numerator, ie:

Assign x to j with largest $e^{-\frac{1}{2}\left(\frac{x-\mu_j}{\sigma}\right)^2} \pi_j$, $j = 1, \dots, K$

or equivalently:

Assign x to j with largest $\log \pi_j - \frac{1}{2}\left(\frac{x-\mu_j}{\sigma}\right)^2 =$

$$= \log \pi_j - \frac{x^2}{2\sigma^2} + \frac{x\mu_j}{\sigma^2} - \frac{\mu_j^2}{2\sigma^2}$$

which is finally equivalent to

Assign x to class j with largest

$$\delta_j(x) = \underbrace{\log \pi_j - \frac{\mu_j^2}{2\sigma^2}}_{\beta_0^{(j)}} + \underbrace{\frac{\mu_j}{\sigma^2}}_{\beta_1^{(j)}} x$$

LINEAR IN x FOR EACH j !

LDA FROM DATA

In practice, we can only apply this rule if we know the parameters:

$$\mu_1, \dots, \mu_K, \sigma^2, \pi_1, \dots, \pi_K$$

LDA approximates Bayes rule (under a Gaussian assumption) by plugging in the ML estimates of the parameters:

$$\hat{\mu}_j = \sum_{i: y_i = j} \frac{x_i}{n_j} \qquad \text{(sample mean in class } j\text{)}$$

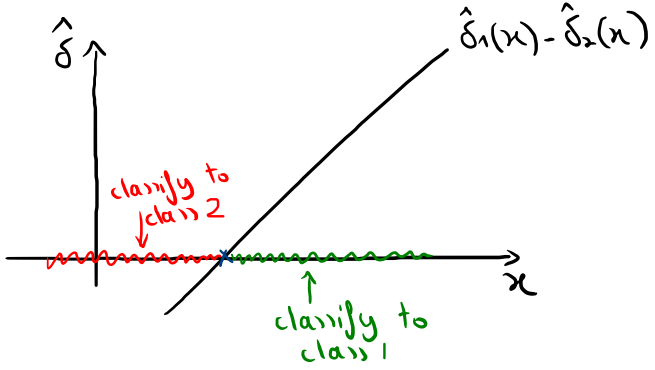
$$\sigma^2 = \frac{\sum_{j=1}^{K} (n_j - 1) s_j^2}{\sum_{j=1}^{K} (n_j - 1)} = \frac{\sum_{j=1}^{K} \sum_{i: y_i = j} (x_i - \hat{\mu}_j)^2}{n - K}$$

(weighted average of class variances or pooled variance)

Then

$$\hat{\delta}_j(x) = \log(\hat{\pi}_j) - \frac{\hat{\mu}_j^2}{2\hat{\sigma}^2} + x \frac{\hat{\mu}_j}{\hat{\sigma}^2}, \quad j = 1, \dots, K$$

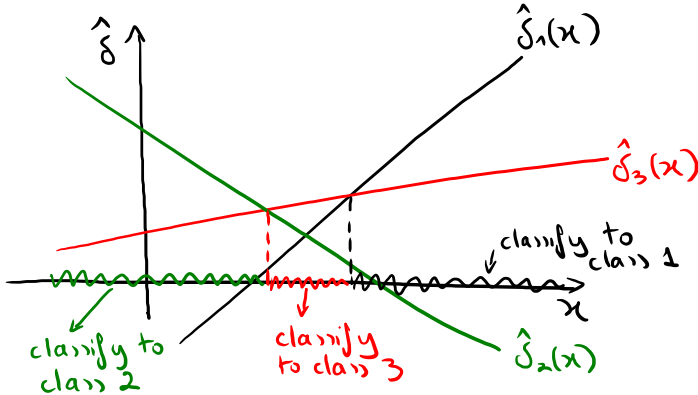
SIMPLE VISUALIZATION ($p = 1$)



$K = 2$

Assign x to class 1 if

$$\hat{\delta}_1(x) > \hat{\delta}_2(x)$$



$K = 3$

Assign x to class j with largest $\hat{\delta}_j(x)$

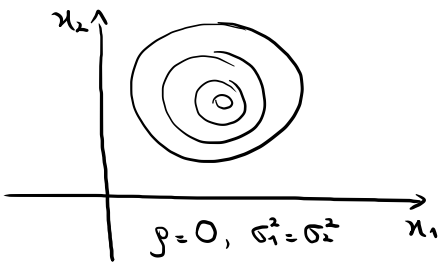
MULTIVARIATE LDA ($p>1$)

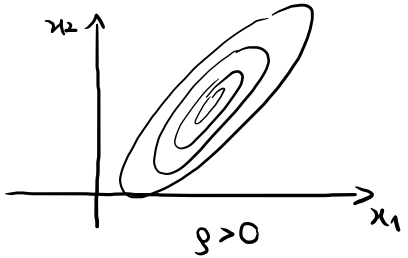
In this case $X = (X_1, \dots, X_p)$ and we assume that it is drawn from a multivariate Gaussian distribution with a class-specific mean vector and a covariance matrix that is common across classes, ie

$$X \mid Y = j \sim N_p(\underline{\mu}_j, \Sigma), \text{ ie } \Sigma_1 = \Sigma_2 = \dots = \Sigma_K = \Sigma$$

$$f_j(\underline{x}) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\underline{x} - \underline{\mu}_j)^t \Sigma^{-1} (\underline{x} - \underline{\mu}_j)\right\}$$

with $\underline{\mu}_j = \begin{pmatrix} \mu_{1j} \\ \vdots \\ \mu_{pj} \end{pmatrix}$ and $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \vdots & & \ddots & \\ \sigma_{p1} & & & \sigma_p^2 \end{pmatrix}$





MULTIVARIATE LDA FOR CLASSIFICATION

Using Bayes theorem with the new $f_j(\underline{x})$ leads to

$$P(Y=j \mid X=\underline{x}) = \frac{\pi_j f_j(x)}{\sum_{i=1}^{k} \pi_i f_i(x)}$$

$$= \frac{\pi_j \frac{1}{(2\pi)^{p/2}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\underline{x}-\underline{\mu}_j)^t \Sigma^{-1}(\underline{x}-\underline{\mu}_j)\right\}}{\sum_{i=1}^{k} \pi_i \frac{1}{(2\pi)^{p/2}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\underline{x}-\underline{\mu}_i)^t \Sigma^{-1}(\underline{x}-\underline{\mu}_i)\right\}}$$

As before, this leads to maximising

$$\delta_j(\underline{x}) = \log(\pi_j) - \frac{1}{2}(\underline{x}-\underline{\mu}_j)^t \Sigma^{-1}(\underline{x}-\underline{\mu}_j) \qquad j = 1, \ldots, K$$

$$= \log(\pi_j) - \underbrace{\frac{1}{2}\underline{x}^t \Sigma^{-1}\underline{x}}_{\text{constant}} + \underline{x}^t \Sigma^{-1}\underline{\mu}_j - \frac{1}{2}\underline{\mu}_j^t \Sigma^{-1}\underline{\mu}_j$$

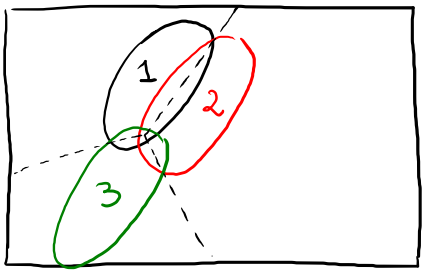
$$\propto \underbrace{\log(\pi_j) - \frac{1}{2}\underline{\mu}_j^t \Sigma^{-1}\underline{\mu}_j}_{\beta_0^{(j)}} + \underbrace{\underline{x}^t \Sigma^{-1}\underline{\mu}_j}_{\underline{\beta}^{(j)} = \begin{pmatrix} \beta_1^{(j)} \\ \vdots \\ \beta_p^{(j)} \end{pmatrix}} = \beta_0^{(j)} + \beta_1^{(j)} x_1 + \cdots + \beta_p^{(j)} x_p$$

LINEAR!

LDA: LINEAR DECISION SURFACES

Decision surface is also linear, as for any 2 classes i and j, this is given by

$$\underline{x}: \quad \delta_i(\underline{x}) = \delta_j(\underline{x}) \quad \text{, ie a } p\text{-}1 \text{ dimensional hyperplane}$$



As before, the parameters of $f_j(\underline{x})$ are unknown and estimated from data:

$$\hat{\pi}_j = \frac{n_j}{n}$$

$$\hat{\mu}_j = \frac{1}{n_j} \sum_{i: y_i = j} \underline{x}_i \quad \text{(vector of sample means)}$$

$$\hat{\Sigma} = \frac{1}{n-K} \sum_{j=1}^{K} \sum_{i: y_i = j} (\underline{x}_i - \hat{\mu}_j)(\underline{x}_i - \hat{\mu}_j)^t \quad \text{(pooled sample covariance)}$$

Leading to $\hat{\delta}_j(\underline{x})$ and to the probabilities

$$\widehat{P(y = j \mid \underline{x})} = \frac{e^{\hat{\delta}_j(\underline{x})}}{\sum_{i=1}^{K} e^{\hat{\delta}_i(\underline{x})}}$$

on which we can set a threshold for classification.

EXTENSION: QUADRATIC DISCRIMINANT ANALYSIS (QDA)

Let's go back to the main assumptions in LDA:

$$X \mid Y = j \sim N_p(\underline{\mu}_j, \Sigma) \qquad \text{LDA}$$

We now drop the second assumption (of equal covariance across classes):

$$X \mid Y = j \sim N_p(\underline{\mu}_j, \Sigma_j) \qquad \text{QDA}$$

This allows more flexibility to the expense of many more parameters to estimate.

How do the decision boundaries look like for multivariate QDA?

QDA: QUADRATIC DECISION SURFACES

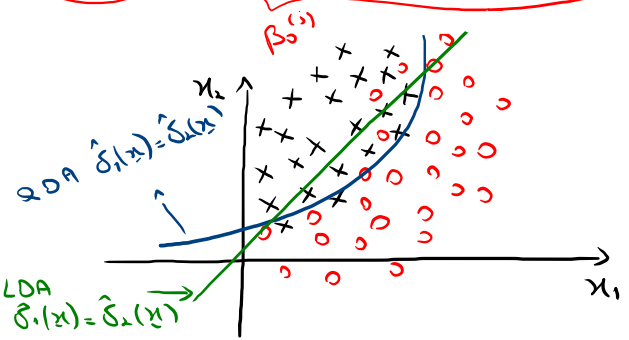
Using Bayes theorem with the new $\delta_j(\underline{x})$:

$$P(Y=j \mid X=\underline{x}) = \frac{\pi_j \frac{1}{(2\pi)^{p/2}|\Sigma_j|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\underline{x}-\underline{\mu}_j)^t \Sigma_j^{-1}(\underline{x}-\underline{\mu}_j)\right\}}{\sum_{i=1}^{K} \pi_i \frac{1}{(2\pi)^{p/2}|\Sigma_i|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\underline{x}-\underline{\mu}_i)^t \Sigma_i^{-1}(\underline{x}-\underline{\mu}_i)\right\}}$$

Leading to:

$$\begin{aligned}\delta_j(\underline{x}) &= \log(\pi_j) - \frac{1}{2}\log|\Sigma_j| - \frac{1}{2}(\underline{x}-\underline{\mu}_j)^t \Sigma_j^{-1}(\underline{x}-\underline{\mu}_j) \\ &= \log(\pi_j) - \frac{1}{2}\log|\Sigma_j| - \frac{1}{2}\underline{x}^t\Sigma_j^{-1}\underline{x} + \underline{x}^t\Sigma_j^{-1}\underline{\mu}_j - \frac{1}{2}\underline{\mu}_j^t\Sigma_j^{-1}\underline{\mu}_j \\ &= \log(\pi_j) - \frac{1}{2}\log|\Sigma_j| - \frac{1}{2}\underline{\mu}_j^t\Sigma_j^{-1}\underline{\mu}_j + \underline{x}^t\Sigma_j^{-1}\underline{\mu}_j - \frac{1}{2}\underline{x}^t\Sigma_j^{-1}\underline{x}\end{aligned}$$

$\beta_0^{(j)}$ (first group) $\quad \underline{x}^t\underline{\beta}^{(j)}$ (second group) $\quad$ quadratic term!



Similar to before:

$$\hat{\underline{\mu}}_j = \frac{\sum_{i:y_i=j} \underline{x}_i}{n_j}$$

$$\hat{\Sigma}_j = \frac{\sum_{i:y_i=j} (\underline{x}_i-\hat{\underline{\mu}}_j)(\underline{x}_i-\hat{\underline{\mu}}_j)^t}{n_j-1}$$

FINAL COMMENTS

① QDA vs LDA

QDA more complex than LDA, hence lower bias but higher variance

parameters: QDA: $K \cdot p + K\,p(p+1)/2$ e.g $K=2$, $p=50$ QDA: 2650 parameters

LDA: $Kp + p(p+1)/2$ LDA: 1375 parameters

If covariances are very different to each other, than LDA will not do well. A choice can be made on unseen / test data

② LDA vs Logistic Regression

For simplicity, take $K=2$ and $p=1$, then

$$\text{log odds}_{LDA} = \log\left(\frac{p(1|x)}{1-p(1|x)}\right) = \delta_1(x) - \delta_0(x) = \alpha_0 + \alpha_1 x$$

$$\text{log odds}_{LR} = \log\left(\frac{p(1|x)}{1-p(1|x)}\right) = \beta_0 + \beta_1 x$$

FINAL COMMENTS (Ctd)

Both LDA and LR have linear decision surfaces. The only difference is in the estimation procedure: in LR, β_0 and β_1 are estimated by ML directly, whereas in LDA, α_0 and α_1 depend on means and covariances of $f_j(x)$ which are estimated by ML under a Gaussian assumption. Whether LDA is better than LR depends on how realistic this assumption is.

(3) K-nn vs LDA and LR

K-nn is a non-parametric approach, so it will work better when the decision surface is highly non-linear and QDA is not enough. Of course, K-nn is not suited for inference.