S&DS 242/542: Theory of Statistics

Lecture 25: Generative models for classification

Review: Logistic regression

Consider predictors $X = (x_1, \ldots, x_p) \in \mathbb{R}^p$ and a classification task with binary response $Y \in \{0, 1\}$.

In logistic regression, we model

$$p(X) = \mathbb{P}[Y = 1 \mid X] = \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}$$

Equivalently, the log-odds ratio of $Y = 1$ to $Y = 0$ is modeled as

$$\log \frac{p(X)}{1 - p(X)} = \log \frac{\mathbb{P}[Y = 1 \mid X]}{\mathbb{P}[Y = 0 \mid X]} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

This approach makes minimal assumptions about the distribution of X, and directly models the distribution of Y given X.

Generative classification models

An alternative approach to classification is to model the distribution of X in each class, i.e. the distribution of X given Y.

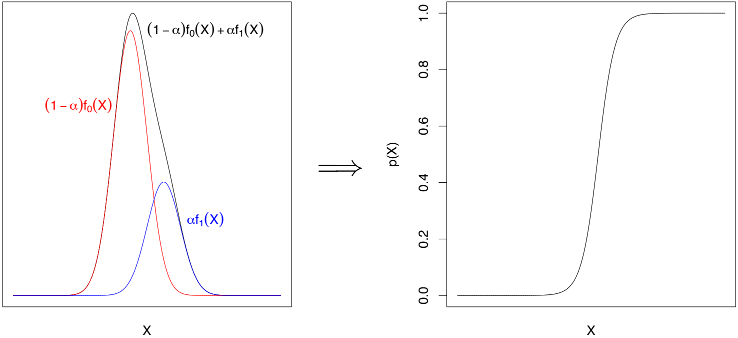
Let $f_0(x)$ and $f_1(x)$ be the PDFs or PMFs of X in classes 0 and 1, and let $\alpha = \mathbb{P}[Y = 1]$ be the marginal probability of class 1.

Generative models for classification then estimate $\mathbb{P}[Y = 1 \mid X]$ by learning $f_0(x)$, $f_1(x)$, and α, and applying Bayes' rule

$$p(X) = \mathbb{P}[Y = 1 \mid X] = \frac{\alpha f_1(X)}{(1-\alpha) f_0(X) + \alpha f_1(X)}$$

Methods differ in how they model and estimate $f_0(x)$ and $f_1(x)$.

Generative classification models



This approach makes stronger assumptions about the distribution of X. It can yield more accurate estimates of $p(X)$ if these assumptions are correct and the distribution of X within each class can be accurately estimated.

Simple linear discriminant analysis

Linear discriminant analysis

Consider a single predictor $X \in \mathbb{R}$ and binary response $Y \in \{0, 1\}$. **Linear discriminant analysis** assumes that:

- $\mathbb{P}[Y = 1] = \alpha$ and $\mathbb{P}[Y = 0] = 1 - \alpha$
- X has a normal distribution in each class $Y = 0$ and $Y = 1$
- The variance of X is the same in both classes

Thus

$$f_0(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_0)^2}{2\sigma^2}} \quad \text{and} \quad f_1(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}}$$

This model for (X, Y) has four unknown parameters $\alpha, \mu_0, \mu_1, \sigma^2$.

Log-odds ratio for LDA

$$\mathbb{P}[Y=1|X] = \frac{\alpha f_1(X)}{(1-\alpha) f_0(X) + \alpha f_1(X)}, \quad \mathbb{P}[Y=0|X] = \frac{(1-\alpha) f_0(X)}{(1-\alpha) f_0(X) + \alpha f_1(X)}$$

$$\Rightarrow \frac{\mathbb{P}[Y=1|X]}{\mathbb{P}[Y=0|X]} = \frac{\alpha f_1(X)}{(1-\alpha) f_0(X)} = \frac{\alpha \cdot e^{-\frac{(X-\mu_1)^2}{2\sigma^2}}}{(1-\alpha) \cdot e^{-\frac{(X-\mu_0)^2}{2\sigma^2}}}$$

$$\Rightarrow \log \frac{\mathbb{P}[Y=1|X]}{\mathbb{P}[Y=0|X]} = \log \frac{\alpha}{1-\alpha} - \frac{(X-\mu_1)^2}{2\sigma^2} + \frac{(X-\mu_0)^2}{2\sigma^2}$$

$$= \log \frac{\alpha}{1-\alpha} - \frac{X^2 - 2\mu_1 X + \mu_1^2}{2\sigma^2} + \frac{X^2 - 2\mu_0 X + \mu_0^2}{2\sigma^2}$$

$$= \log \frac{\alpha}{1-\alpha} + \frac{\mu_1 - \mu_0}{\sigma^2} \cdot X - \frac{\mu_1^2}{2\sigma^2} + \frac{\mu_0^2}{2\sigma^2}$$

Log-odds ratio for LDA

Thus, under the LDA model,

$$\log \frac{\mathbb{P}[Y = 1 \mid X]}{\mathbb{P}[Y = 0 \mid X]} = \underbrace{\log \frac{\alpha}{1-\alpha} - \frac{\mu_1^2}{2\sigma^2} + \frac{\mu_0^2}{2\sigma^2}}_{=\beta_0} + \underbrace{\frac{\mu_1 - \mu_0}{\sigma^2}}_{=\beta_1} X$$

Equivalently,

$$\mathbb{P}[Y = 1 \mid X] = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

This has the same form as logistic regression, where the log-odds ratio of $Y = 1$ to $Y = 0$ is linear in X.

However, LDA estimates β_0, β_1 in a different way, by estimating the parameters $\alpha, \mu_0, \mu_1, \sigma^2$ and computing β_0, β_1 from these estimates.

Maximum likelihood estimation

Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are IID and distributed according to the LDA model. To compute the log-likelihood:

$$f(X,Y) = \left(\alpha \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu_1)^2}{2\sigma^2}}\right)^{\mathbb{1}\{Y=1\}} \cdot \left((1-\alpha) \cdot \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu_0)^2}{2\sigma^2}}\right)^{\mathbb{1}\{Y=0\}}$$

$$\Rightarrow \ell_n(\alpha, \mu_0, \mu_1, \sigma^2) = \sum_{i=1}^{n} \log f(X_i, Y_i)$$

$$= \sum_{i=1}^{n} \Big[\mathbb{1}\{Y_i = 1\} \left(\log \alpha - \frac{1}{2} \log 2\pi\sigma^2 - \frac{(X_i - \mu_1)^2}{2\sigma^2} \right)$$

$$+ \mathbb{1}\{Y_i = 0\} \left(\log (1-\alpha) - \frac{1}{2} \log 2\pi\sigma^2 - \frac{(X_i - \mu_0)^2}{2\sigma^2} \right) \Big]$$

$$= \underbrace{N_1}_{\#\, Y_i\text{'s} = 1} \cdot \log \alpha + \underbrace{N_0}_{\#\, Y_i\text{'s} = 0} \cdot \log (1-\alpha) - \frac{n}{2} \log 2\pi\sigma^2 - \sum_{i : Y_i = 1} \frac{(X_i - \mu_1)^2}{2\sigma^2} - \sum_{i : Y_i = 0} \frac{(X_i - \mu_0)^2}{2\sigma^2}$$

Maximum likelihood estimation

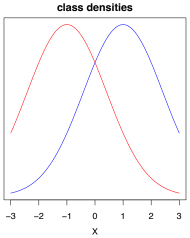
Thus the log-likelihood is

$$\ell_n(\alpha, \mu_0, \mu_1, \sigma^2) = N_1 \log \alpha + N_0 \log(1-\alpha) - \frac{n}{2} \log 2\pi\sigma^2 \\ - \sum_{i:Y_i=1} \frac{(X_i - \mu_1)^2}{2\sigma^2} - \sum_{i:Y_i=0} \frac{(X_i - \mu_0)^2}{2\sigma^2}$$

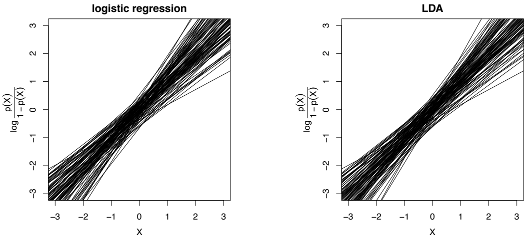
The MLEs are given by:

- $\hat{\alpha} = \frac{N_1}{N_0+N_1} = \frac{N_1}{n}$, the sample proportion of class 1.
- $\hat{\mu}_1 = \frac{1}{N_1} \sum_{i:Y_i=1} X_i$ and $\hat{\mu}_0 = \frac{1}{N_0} \sum_{i:Y_i=0} X_i$, the sample average within each class.
- $\hat{\sigma}^2 = \frac{1}{n}[\sum_{i:Y_i=1}(X_i - \hat{\mu}_1)^2 + \sum_{i:Y_i=0}(X_i - \hat{\mu}_0)^2]$. For small sample sizes n, it is also common to use instead the unbiased estimator with normalization $\frac{1}{n-2}$ instead of $\frac{1}{n}$.

LDA vs. logistic regression

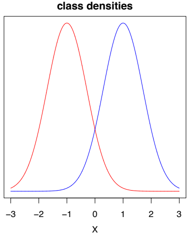


Class 0: $\mathcal{N}(-1, 2)$, Class 1: $\mathcal{N}(1, 2)$ (with $\alpha = 0.5$)

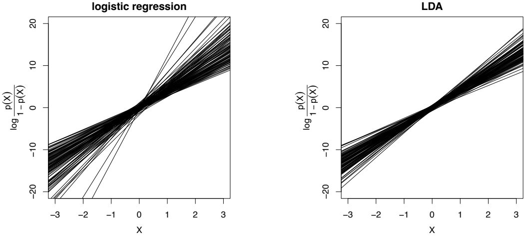


Estimated log-odds for sample size $n = 100$, across 100 trials

LDA vs. logistic regression



Class 0: $\mathcal{N}(-1, 0.5)$, Class 1: $\mathcal{N}(1, 0.5)$ (with $\alpha = 0.5$)



Estimated log-odds for sample size $n = 100$, across 100 trials

Fisher information matrix

$$\log f(X,Y|\alpha,\mu_0,\mu_1,\sigma^2) = \mathbb{I}\{Y=1\}\cdot\log\alpha + \mathbb{I}\{Y=0\}\cdot\log(1-\alpha) - \frac{1}{2}\log 2\pi\sigma^2$$
$$- \mathbb{I}\{Y=1\}\cdot\frac{(X-\mu_1)^2}{2\sigma^2} - \mathbb{I}\{Y=0\}\cdot\frac{(X-\mu_0)^2}{2\sigma^2}$$

$$\frac{\partial}{\partial\alpha}\log f(X,Y|\alpha,\mu_0,\mu_1,\sigma^2) = \frac{\mathbb{I}\{Y=1\}}{\alpha} - \frac{\mathbb{I}\{Y=0\}}{1-\alpha}$$

$$\mathbb{E}\left[\frac{\partial^2}{\partial\alpha^2}\log f(X,Y|\alpha,\mu_0,\mu_1,\sigma^2)\right] = \mathbb{E}\left[-\frac{\mathbb{I}\{Y=1\}}{\alpha^2} - \frac{\mathbb{I}\{Y=0\}}{(1-\alpha)^2}\right]$$
$$= -\frac{\alpha}{\alpha^2} - \frac{(1-\alpha)}{(1-\alpha)^2} = -\frac{1}{\alpha} - \frac{1}{1-\alpha} = -\frac{1}{\alpha(1-\alpha)}$$

$$\frac{\partial}{\partial\mu_0}\log f(X,Y|\alpha,\mu_0,\mu_1,\sigma^2) = \mathbb{I}\{Y=0\}\frac{X-\mu_0}{\sigma^2}$$

$$\mathbb{E}\left[\frac{\partial^2}{\partial\mu_0^2}\log f(X,Y|\alpha,\mu_0,\mu_1,\sigma^2)\right] = \mathbb{E}\left[\mathbb{I}\{Y=0\}\cdot\frac{-1}{\sigma^2}\right] = -\frac{1-\alpha}{\sigma^2}$$

Fisher information matrix

$$\mathbb{E}\left[\frac{\partial^2}{\partial\mu_0\partial\sigma^2}\log f(X,Y|\alpha,\mu_0,\mu_1,\sigma^2)\right] = \mathbb{E}\left[-\mathbb{I}\{Y=0\}\cdot\frac{X-\mu_0}{\sigma^4}\right]$$

$$= -(1-\alpha)\cdot\mathbb{E}\left[\frac{X-\mu_0}{\sigma^4}\Big|Y=0\right] = 0$$

Similarly $\mathbb{E}\left[\frac{\partial^2}{\partial\mu_1^2}\log f(\ldots)\right] = -\frac{\alpha}{\sigma^2}$, $\mathbb{E}\left[\frac{\partial^2}{\partial\mu_1\partial\sigma^2}\log f(\ldots)\right] = 0$

$$\frac{\partial}{\partial\sigma^2}\log f(X,Y|\alpha,\mu_0,\mu_1,\sigma^2) = -\frac{1}{2\sigma^2} + \mathbb{I}\{Y=1\}\cdot\frac{(X-\mu_1)^2}{2(\sigma^2)^2} + \mathbb{I}\{Y=0\}\cdot\frac{(X-\mu_0)^2}{2(\sigma^2)^2}$$

$$\mathbb{E}\left[\frac{\partial^2}{\partial(\sigma^2)^2}\log f(X,Y|\alpha,\mu_0,\mu_1,\sigma^2)\right] = \mathbb{E}\left[\frac{1}{2\sigma^4} - \mathbb{I}\{Y=1\}\cdot\frac{(X-\mu_1)^2}{(\sigma^2)^3} - \mathbb{I}\{Y=0\}\cdot\frac{(X-\mu_0)^2}{(\sigma^2)^3}\right]$$

$$= \frac{1}{2\sigma^4} - \alpha\cdot\mathbb{E}\left[\frac{(X-\mu_1)^2}{(\sigma^2)^3}\Big|Y=1\right] - (1-\alpha)\cdot\mathbb{E}\left[\frac{(X-\mu_0)^2}{(\sigma^2)^3}\Big|Y=0\right]$$

$$= \frac{1}{2\sigma^4} - \frac{\alpha}{\sigma^4} - \frac{1-\alpha}{\sigma^4} = -\frac{1}{2\sigma^4}$$

Standard errors for parameter estimates

The Fisher information matrix is the diagonal matrix

$$I(\alpha, \mu_0, \mu_1, \sigma^2) = \begin{pmatrix} \frac{1}{\alpha(1-\alpha)} & & & \\ & \frac{1-\alpha}{\sigma^2} & & \\ & & \frac{\alpha}{\sigma^2} & \\ & & & \frac{1}{2\sigma^4} \end{pmatrix}$$

Thus for large n, the distributions of the model parameter MLEs are approximately

$$\hat{\alpha} \sim \mathcal{N}(\alpha, \tfrac{\alpha(1-\alpha)}{n}), \qquad \hat{\mu}_0 \sim \mathcal{N}(\mu_0, \tfrac{\sigma^2}{n(1-\alpha)})$$

$$\hat{\mu}_1 \sim \mathcal{N}(\mu_1, \tfrac{\sigma^2}{n\alpha}), \qquad \hat{\sigma}^2 \sim \mathcal{N}(\sigma^2, \tfrac{2\sigma^4}{n}),$$

and these MLEs are approximately uncorrelated with each other.

Standard error of $\hat{\beta}_1$

Recalling $\beta_1 = (\mu_1 - \mu_0)/\sigma^2$, and applying the preceding forms of the standard errors for μ_0, μ_1, σ^2 and the delta method, we have

$$\hat{\beta}_1 = \frac{\hat{\mu}_1 - \hat{\mu}_0}{\hat{\sigma}^2} \sim \mathcal{N}\Big(\beta_1,\ \frac{1}{n\sigma^2\alpha(1-\alpha)} + \frac{2(\mu_1 - \mu_0)^2}{n\sigma^4}\Big)\Big)$$

approximately for large n. Fixing σ^2 and α, the variance of $\hat{\beta}_1$ grows linearly in $(\mu_1 - \mu_0)^2$.

For logistic regression, recall $\hat{\beta}_1 \sim \mathcal{N}(\beta_1, \frac{1}{n}[I(\beta)^{-1}]_{11})$ where

$$I(\beta) = \mathbb{E}\left[p(X \mid \beta)(1 - p(X \mid \beta)) \begin{pmatrix} 1 & X \\ X & X^2 \end{pmatrix}\right]$$

and $[I(\beta)^{-1}]_{11}$ is the lower-right entry. It may be shown that this variance of $\hat{\beta}_1$ grows exponentially in $(\mu_1 - \mu_0)^2$, and thus is much larger than that of LDA when the class means are well-separated.

Generative classification models

LDA with multiple predictors and classes

More generally, for predictors $X \in \mathbb{R}^p$ and K classes $Y \in \{0, 1, \ldots, K-1\}$, **linear discriminant analysis** assumes that:

- $\mathbb{P}[Y = k] = \alpha_k$ for each $k = 0, 1, \ldots, K-1$
- X has a multivariate normal distribution in $\mathbb{R}^p$ for each class
- The covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$ of this multivariate normal distribution is the same in all classes

Thus

$$f_k(x) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} e^{-\frac{1}{2}(x-\mu_k)^\top \Sigma^{-1}(x-\mu_k)} \text{ for } k = 0, 1, \ldots, K-1$$

The parameters are the class probabilities $\alpha = (\alpha_0, \ldots, \alpha_{K-1})$, class means $\mu_0, \ldots, \mu_{K-1} \in \mathbb{R}^p$, and class covariance $\Sigma \in \mathbb{R}^{p \times p}$.

Log-odds ratios for LDA

$$\log \frac{\mathbb{P}[Y=k|X]}{\mathbb{P}[Y=0|X]} = \log \frac{\alpha_k \cdot f_k(x)}{\alpha_0 \cdot f_0(x)}$$

$$= \log \frac{\alpha_k}{\alpha_0} + \log \frac{e^{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)}}{e^{-\frac{1}{2}(x-\mu_0)^T \Sigma^{-1}(x-\mu_0)}}$$

$$= \log \frac{\alpha_k}{\alpha_0} - \frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k) + \frac{1}{2}(x-\mu_0)^T \Sigma^{-1}(x-\mu_0)$$

$$= \log \frac{\alpha_k}{\alpha_0} + (\mu_k-\mu_0)^T \Sigma^{-1} x - \frac{1}{2}\mu_k^T \Sigma^{-1}\mu_k + \frac{1}{2}\mu_0^T \Sigma^{-1}\mu_0$$

Log-odds ratios for LDA

Thus, under the LDA model, for $X = (x_1, \ldots, x_p)$,

$$\log \frac{\mathbb{P}[Y = k \mid X]}{\mathbb{P}[Y = 0 \mid X]} = \underbrace{\log \frac{\alpha_k}{\alpha_0} - \frac{1}{2}\mu_k \Sigma^{-1} \mu_k + \frac{1}{2}\mu_0 \Sigma^{-1} \mu_0}_{=\beta_{k0}}$$

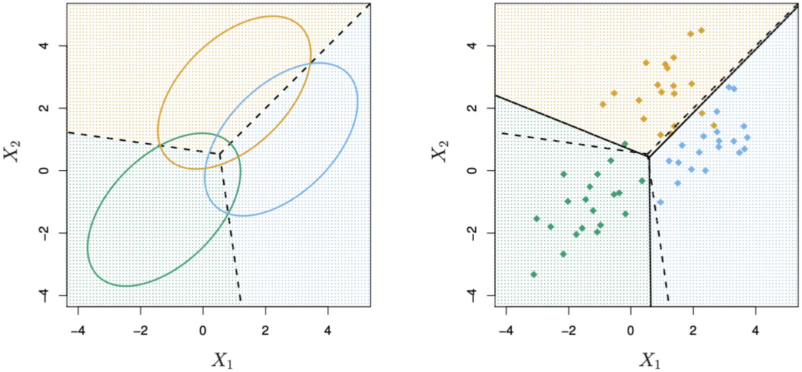
$$+ \underbrace{[\Sigma^{-1}(\mu_k - \mu_0)]_1}_{=\beta_{k1}} x_1 + \ldots + \underbrace{[\Sigma^{-1}(\mu_k - \mu_0)]_p}_{=\beta_{kp}} x_p$$

This has the same form as multinomial logistic regression, but $\beta_{k0}, \ldots, \beta_{kp}$ are estimated via the MLEs

$$\hat{\alpha}_k = \frac{N_k}{n}, \qquad \hat{\mu}_k = \frac{1}{N_k} \sum_{i : Y_i = k} X_i,$$

$$\hat{\Sigma} = \frac{1}{n} \sum_{k=0}^{K-1} \sum_{i : Y_i = k} (X_i - \hat{\mu}_k)(X_i - \hat{\mu}_k)^\top$$

LDA with multiple predictors and classes



Suppose we predict $\hat{Y}(X) = k$ if class k has the highest probability $\mathbb{P}[Y = k \mid X]$. (This minimizes the error $\mathbb{P}[Y \neq \hat{Y}(X)]$.)

Then $\hat{Y}(X) = k$ if $\log \frac{\mathbb{P}[Y=k|X]}{\mathbb{P}[Y=\ell|X]} > 0$ for all $\ell \neq k$, so the decision boundaries of LDA are *linear* functions of x.

Quadratic discriminant analysis

Quadratic discriminant analysis extends LDA by assuming that the distribution of X within each class k is

$$X \sim \mathcal{N}(\mu_k, \Sigma_k)$$

with a possibly different covariance $\Sigma_k \in \mathbb{R}^{p \times p}$ for each class.

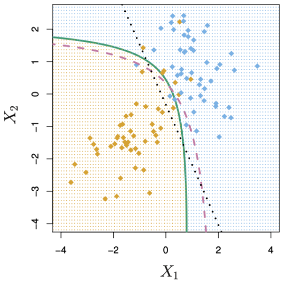
The model parameters may be estimated via the MLEs

$$\hat{\alpha}_k = \frac{N_k}{n}, \qquad \hat{\mu}_k = \frac{1}{N_k} \sum_{i:Y_i=k} X_i,$$

$$\hat{\Sigma}_k = \frac{1}{N_k} \sum_{i:Y_i=k} (X_i - \hat{\mu}_k)(X_i - \hat{\mu}_k)^\top$$

where the covariance Σ_k is now estimated separately for each class.

Log-odds ratios for QDA



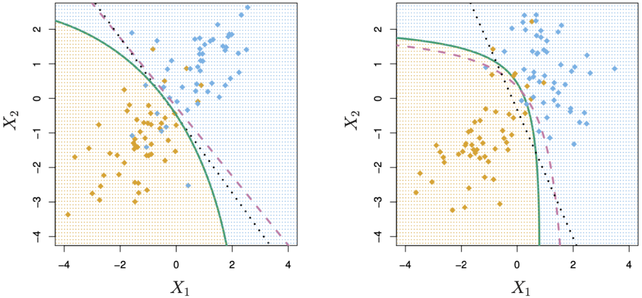
The log-odds ratios

$$\log \frac{\mathbb{P}[Y = k \mid X]}{\mathbb{P}[Y = 0 \mid X]} = \log \frac{\alpha_k}{\alpha_0} - \frac{1}{2} \log \frac{\det \Sigma_k}{\det \Sigma_0} - \frac{1}{2}(X - \mu_k)^\top \Sigma_k^{-1}(X - \mu_k) + \frac{1}{2}(X - \mu_0)^\top \Sigma_0^{-1}(X - \mu_0)$$

and class decision boundaries are *quadratic* functions of X.

Bias-variance tradeoffs

Decision boundaries for LDA and QDA



Left: Data from two classes with common covariance
Right: Data from two classes with differing covariances

LDA is a special case of QDA, with fewer model parameters. Its predictions have smaller variance than QDA, but larger bias when the variance-covariance matrices of the classes are not the same.

Naive Bayes

Estimating a general distribution for $X = (x_1, \ldots, x_p) \in \mathbb{R}^p$ may be challenging if p is large. LDA and QDA achieve this by making the strong assumption that $f_0(X), \ldots, f_{K-1}(X)$ are normal.

A popular alternative called **naive Bayes** instead assumes that the distribution of the p predictors in each class are independent, i.e.

$$f_k(X) = f_{k1}(x_1) \times \ldots \times f_{kp}(x_p)$$

for some univariate distributions $f_{k1}(x_1), \ldots, f_{kp}(x_p)$. Thus it models only the marginal distribution of each predictor within each class, and not their joint distribution.

The distributions $f_{kj}(x_j)$ do not need to be normal. They are sometimes modeled and estimated assuming a parametric model, and sometimes estimated nonparametrically.

Log-odds ratios for naive Bayes

The log-odds ratio of $Y = k$ to $Y = 0$ takes the form

$$\log \frac{\mathbb{P}[Y = k \mid X]}{\mathbb{P}[Y = 0 \mid X]} = \log \frac{\alpha_k f_{k1}(x_1) \ldots f_{kp}(x_p)}{\alpha_0 f_{01}(x_1) \ldots f_{0p}(x_p)}$$
$$= \underbrace{\log \frac{\alpha_k}{\alpha_0}}_{=\beta_{k0}} + \sum_{j=1}^{p} \underbrace{\log \frac{f_{kj}(x_j)}{f_{0j}(x_j)}}_{=g_{kj}(x_j)}$$

If the distributions $f_{kj}(x_j)$ are completely general, then $g_{kj}(x_j)$ is also a general function of x_j. Thus the log-odds ratio is an *additive* function of $x_1, \ldots, x_p$ which may be nonlinear in each predictor.

If $f_{kj}(x_j)$ is the $\mathcal{N}(\mu_{kj}, \sigma_j^2)$ density, then $g_{kj}(x) = \frac{\mu_{kj} - \mu_{0j}}{\sigma_j^2} \cdot x$ and this becomes a special case of LDA where Σ is diagonal.