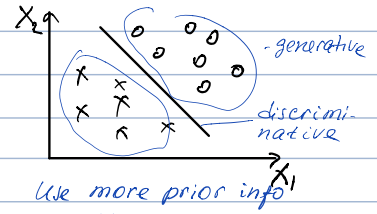
Generative models



Use more prior info
⇒ often cheaper to compute

Bayes rule:

$$P\{Y=k \mid X\} = \frac{f_k(x)\cdot \pi_k}{\sum_{\ell=1}^{K} f_\ell(x)\pi_\ell} = \frac{P\{X \mid Y=k\}\,P\{Y=k\}}{P\{X\}}$$

- $f_k(x)$: distribution of X in a class → likelihood
- π_k: class prior
- denominator: normalizing constant, does not depend on k
- $P\{Y=k \mid X\}$: most probable class, given the data

The MAP estimator minimizing Bayes risk under 0-1 loss is

$$\hat{Y} = \arg\max_k P\{Y=k \mid X\} = \arg\max_k f_k(X)\pi_k$$ — ignore the denominator

Specification of $f_k(x)$ define the classifier:

- $f_k(x)$ multivariate Gaussian, same Σ (variance-covariance) per class → LDA (Linear discriminant analysis)
- $f_k(x)$ multivariate Gaussian, different Σ_k → QDA (quadratic discriminant analysis)
- $f(x)$ = mixture of Gaussians; $f_k(x)$ – component of the mixture
- $f_k(X) = \prod_{p=1}^{P} f_k(X_p)$ – naïve Bayes
  - in each class, orthogonal dimensions
  - typically (but not always) $f_k(X_p)$ is non-parametric – see hw3

$$\hat{Y} = \arg\max_k P\{Y=k \mid X\} = \arg\max_k \left[\log f_k(x) + \log \pi_k\right] \overset{MVN}{=} \arg\max_k \left\{-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k) + \log\pi_k\right\}$$

$$= \arg\max_k \left\{-\frac{1}{2}X^T\Sigma^{-1}X + X^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log\pi_k\right\}$$

$\delta_k(X)$ – linear discriminant function

LDA in high dimensions

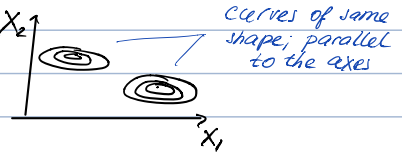
Nearest centroids (diagonal covariance LDA)

When $P \gg N$, estimation of Σ is unstable ⇒ assume indep. predictors

$$X_{i:\,y_i=k} \sim MVN\left(\mu_1, \begin{pmatrix}\sigma_1^2 & & 0\\ & \ddots & \\ 0 & & \sigma_P^2\end{pmatrix}\right); \quad \hat{\mu}_k = \begin{pmatrix}\bar{x}_{\cdot 1}\\ \vdots \\ \bar{x}_{\cdot P}\end{pmatrix}, \quad \hat{\Sigma} = \begin{pmatrix}s_1^2 & & 0\\ & \ddots & \\ 0 & & s_P^2\end{pmatrix}$$

over all $\{i: y_i=k\}$

pooled variance across classes: $s_j^2 = \frac{1}{N-K}\sum_{k=1}^{K}\sum_{\{i:y_i=k\}}(x_{ij}-\bar{x}_{kj})^2$



$$\hat{y}(x_i) = \arg\max_k \left\{-\frac{1}{2}\sum_{j=1}^{P}\frac{(x_{ij}-\mu_j)^2}{\sigma_j^2} + \log\pi_k\right\}$$

Estimation:

$$\hat{y}(x_i) = \arg\min_k \sum_{j=1}^{P}\frac{(x_{ij}-\bar{x}_{\cdot j})^2}{s_j^2} - 2\log\pi_k$$

x_i: new obs; over $\{i: y_i=k\}$

⇒ nearest centroid classifier (after standardization and correction for class prior)

Note: connection to SVD

Connection to Singular Value Decomposition (SVD)

Consider class k. To estimate $\hat{\Sigma}_k$ with class-specific variance-covariance,

(1) center each feature

$X^{(1)}$ centered

$i : y_i = k$

	X_1 ... X_p
1	$X_{11}-\bar{X}_{\cdot 1}$ ... $X_{1p}-\bar{X}_{\cdot p}$
2	$X_{21}-\bar{X}_{\cdot 1}$ ... $X_{2p}-\bar{X}_{\cdot p}$
...	
N_k	$X_{N_k 1}-\bar{X}_{\cdot 1}$ ... $X_{N_k p}-\bar{X}_{\cdot p}$
	$\bar{X}_{\cdot 1}$ ... $\bar{X}_{\cdot p}$

(2) $(N_k - 1)\hat{\Sigma}_k = X'^T X'$

$\begin{pmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \lambda_p \end{pmatrix}$ $\lambda_1 \gg \lambda_2 \gg \dots \gg \lambda_p$

number of $\lambda_j \neq 0$ = rank of X_k

Singular Value Decomposition: $X' = U \Lambda V^T$

diagonal

X': $N_k \times p$; U: $N_k \times p$; Λ: $p \times p$; V^T: $p \times p$

$U^T U = I_{p \times p}$ — U is an orthonormal matrix

$V V^T = I_{p \times p}$ — V is orthonormal

Can write $X'V = U\Lambda$

scaling factor of the projection

projection of columns of X' (i.e., features) onto directions defined by columns of V

Using the SVD decomposition in QDA:

$$\delta_k(X^{new}) = -\frac{1}{2}\underbrace{(X^{new} - \mu_k)^T}_{1 \times p}\underbrace{\Sigma_k^{-1}}_{p \times p}\underbrace{(X^{new} - \mu_k)}_{p \times 1} + \log \pi_k$$

substitute SVD

$$\hat{\Sigma}_k = X_k'^T X' = V \Lambda \underbrace{U^T U}_{I} \Lambda V^T = V \Lambda^2 V^T$$

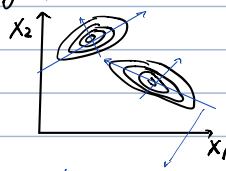
Λ^2

$$\delta_k(X^{new}) = -\frac{1}{2}\left[\underbrace{\Lambda^{-1} V^T (X^{new} - \mu_k)}_{\text{linear transform}}\right]^T I \left[\Lambda^{-1} V^T (X^{new} - \mu_k)\right]$$

level curves are spheres on the transformed scale $\Rightarrow$ classify the new point as the class that has the closest centroid

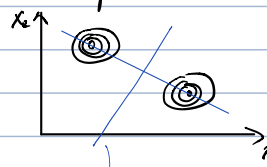
Graphical interpretation:

Original scale:



directions defined by columns of V, scaled by the corresponding λ_i

Transformed scale:



decision boundary on the transformed scale (assuming all π_k equal)

In LDA, same approach; subtract class-specific feature means in (1)

Regularized LDA — Middle ground between LDA and QDA

$$\hat{\Sigma}_k(\alpha) = \alpha \hat{\Sigma}_k + (1-\alpha)\hat{\Sigma}$$

class-specific — pooled

Choose α by cross-validation

Middle ground between LDA and diagonal-covariance LDA

$$\hat{\Sigma}(\alpha) = \alpha \hat{\Sigma} + (1-\alpha)\hat{\sigma}^2 I$$

$p \times p$ identity matrix

Nearest shrunken centroids:

$$\hat{\mu}_k = \begin{pmatrix} \bar{x}_{\cdot 1} \\ \cdots \\ \bar{x}_{\cdot p} \end{pmatrix}$$

$\hat{\mu}_k$: centroid of class k; over all $\{i : y_i = k\}$

$$\hat{\mu} = \begin{pmatrix} \bar{x}_{\cdot 1} \\ \cdots \\ \bar{x}_{\cdot p} \end{pmatrix}$$

$\hat{\mu}$: overall centroid; all i

Define $d_{jk} = \dfrac{\hat{\mu}_{jk} - \hat{\mu}_j}{\sqrt{\frac{1}{n_k} - \frac{1}{N}} \cdot s_j}$

m_k

standardized distance of class centroid from overall centroid in dimension j

Large $d_{jk} \to k$ is a discriminative dimension for class k

$\Rightarrow$

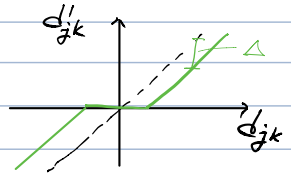
$$\hat{\mu}_{jk} = \hat{\mu}_j + m_k \cdot s_j \cdot d_{jk}$$

when small, dimension j is not informative for class k

Replace with $d'_{jk} = \text{sign}(d_{jk}) \cdot \{ |d_{jk}| - \Delta \}_+$

$$\hat{\mu}'_{jk} = \hat{\mu}_j + m_k \cdot s_j \cdot d'_{jk}$$

choose by cross-validation



soft threshold function

→ remove dimensions that are non-distinguishable from the overall centroid

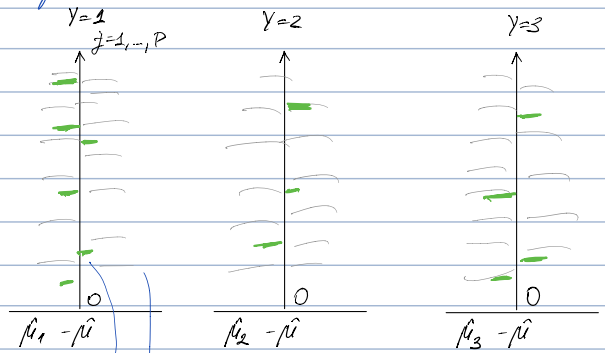
Then the discriminant function is

$$\delta_k(x) = \sum_{j=1}^{p} \frac{x_j - \hat{\mu}'_{jk}}{s_j^2} - 2 \log \pi_k$$

new data point — shrunken centroid

Class probabilities:

$$\hat{\pi}_k(x) = \frac{\exp\{-\frac{1}{2}\delta_k(x)\}}{\sum_{k=1}^{K} \exp\{-\frac{1}{2}\delta_k(x)\}}$$



original centroids

shrunken centroids (smaller; many = 0)

Naive Bayes with categorical predictors:

$$P\{Y=k \mid X\} = \frac{f_k(x)\pi_k}{f(x)} = \frac{\prod_{j=1}^{p} f_k(x_j)\pi_k}{f(x)}$$

estimate empirically (e.g., from histograms or frequencies)

Other uses of Bayesian inference

Discriminative models:

Linear regression: $Y \mid X \sim N(\beta_0 + \beta_1 X, \sigma^2)$

Logistic regression: $Y \mid X \sim \text{Bernoulli}\left(\frac{1}{1+e^{-(\beta_0+\beta_1 X)}}\right)$

↓

Can leverage Bayesian methods in discriminative models. leads to regularization

Generative models

Any Y: $P\{Y \mid X\} = \frac{P\{X \mid Y\} P\{Y\}}{P\{X\}}$

categorical Y: $P\{Y=k \mid X\} = \frac{f_k(x)\pi_k}{\sum_{\ell=1}^{K} f_\ell(x)\pi_\ell}$

apply Bayes rule

Example: Linear regression

$$Y_i = X_i^T \beta_{p\times 1} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \quad \Rightarrow \quad Y_i \sim N(X_i^T\beta, \sigma^2)$$

Frequentist estimation

$$\underbrace{\log L(\beta|X,Y)}_{\text{Log-likelihood}} = \underset{N\times 1 \;\; N\times p \;\; p\times 1}{\text{Log } P\{Y|X,\beta\}} = \sum_{i=1}^N \left[\log \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(Y_i - X_i^T\beta)^2}\right] = \sum_{i=1}^N \left[-\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(Y_i - X_i^T\beta)^2\right]$$

MLE: $\hat\beta_{MLE} = \arg\max_\beta \left[-\frac{1}{2\sigma^2}\sum_{i=1}^N (Y_i - X_i^T\beta)^2\right] = \arg\min_\beta \sum_{i=1}^N (Y_i - X_i^T\beta)^2$ — analytical solution exists

Bayesian estimation

View some quantities of interest as random variables. Here $\beta \sim N(0, \frac{1}{\lambda} I)$ ← assumed prior distr. of β; $p\times p$ identity matrix

$$p(\beta) = \frac{1}{(2\pi)^{p/2}} \cdot e^{-\frac{\lambda}{2}\underbrace{\sum_{j=1}^p \beta_j^2}_{=\beta^T\beta}}$$

← pre-defined constant

$$\underbrace{\log P(\beta|X,Y)}_{\text{Log-posterior prob.}} = \frac{\log P\{Y|X,\beta\}\cdot P\{\beta\}}{P\{Y|X\}}$$

← continue to view X as fixed

MAP: $\hat\beta_{MAP} = \arg\max_\beta P\{\beta|X,Y\} = \arg\max_\beta \left[\underbrace{\log P\{Y|X,\beta\}}_{\text{Log-likelihood}} + \underbrace{\log P(\beta)}_{\text{extra term involves }\beta}\right]$

$$= \arg\max_\beta \left[\sum_{i=1}^N \left\{-\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(Y_i - X_i^T\beta)^2\right\} - \frac{p}{2}\log 2\pi - \frac{\lambda}{2}\beta^T\beta\right]$$

$$= \arg\min_\beta \frac{1}{\sigma^2}\sum_{i=1}^N (Y_i - X_i^T\beta)^2 + \lambda\beta^T\beta = \hat\beta_{Ridge}$$

⇒ Bayesian estimation with priors on parameters β is another motivation for L_2 norm regularization. The prior acts as a regularizer.

Example: Logistic regression

$$Y_i \sim \text{Bernoulli}(\pi_i), \quad \pi_i = \frac{1}{1+e^{-X_i^T\beta}}$$

Frequentist estimation

$$\underbrace{\log L(\beta|X,Y)}_{\text{Log-likelihood}} = \underset{N\times 1 \;\; N\times p \;\; p\times 1}{\text{Log } P\{Y|X,\beta\}} = \sum_{i=1}^N \left[Y_i \log\pi_i + (1-Y_i)\log(1-\pi_i)\right] = \sum_{i=1}^N \left[Y_i \log\frac{1}{1+e^{-X_i^T\beta}} + (1-Y_i)\log\frac{1}{1+e^{X_i^T\beta}}\right]$$

MLE: $\hat\beta_{MLE} = \arg\max_\beta \log L(\beta|X,Y)$ → no analytical solution; gradient descent

Bayesian estimation

$$p(\beta) = \frac{1}{(2\pi)^{p/2}} \cdot e^{-\frac{\lambda}{2}\underbrace{\sum_{j=1}^p \beta_j^2}_{=\beta^T\beta}}$$

MAP: $\hat\beta_{MAP} = \arg\max_\beta P\{\beta|X,Y\} = \arg\max_\beta \left[\underbrace{\log P\{Y|X,\beta\}}_{\text{Log-likelihood}} + \underbrace{\log P(\beta)}_{\text{extra term involves }\beta}\right]$

$$= \arg\max_\beta \left[\log P\{Y|X,\beta\} + \frac{1}{2}\beta^T\beta\right] = \hat\beta_{Ridge}$$

use in gradient descent

Similarly, use Laplace(0,6) prior for L_1 regularization

$$P(\theta) = \frac{1}{2\sigma} e^{-\frac{|\theta - 0|}{\sigma}}$$