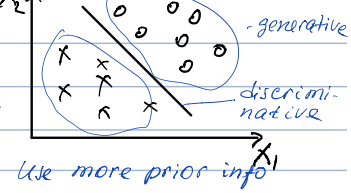
Day 10, 10/9/2018

Generative models



Generative models describe distributions of X in each class

Use more prior info
⇒ often cheaper to compute

Bayes rule: $p(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \propto P(X|Y)P(Y)$

$p(Y|X)$: Linear regression, logistic regression ⇒ discriminative models

$P(X|Y)P(Y)$: generative models

In classification, Y is discrete

$P\{Y=k|X\} = \frac{f_k(X)\cdot\pi_k}{\sum_{l=1}^{K} f_l(X)\pi_l}$

$f_k(X)$: distribution of X in a class → likelihood

π_k: class prior

$\sum_{l=1}^{K} f_l(X)\pi_l$: normalizing constant, does not depend on k

$P\{Y=k|X\}$: most probable class, given the data

Goal:
find class k maximizing the posterior probability
MAP (maximum a posteriori learning)

Data: $\{(x_i, y_i)\}$, $y_i \in \{1,\dots,K\}$ - categorical

More formally: minimize the 0-1 loss function

$J(Y, \hat{Y}(x)) = \begin{cases} 0, & \hat{Y}(x) = Y \text{ — correct prediction} \\ 1, & \hat{Y}(x) \neq Y \text{ — wrong prediction} \end{cases}$

In linear reg: $J(Y,\hat{Y}) = \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$

Bayes risk: loss function, averaged over $f_k(x)$ and k:

$R(\hat{Y}(x)) = E\{E\{J(Y,\hat{Y}(x))\}\}$

over $f(X|Y=k)$ - all data of a class; over π_k - all classes

In linear regression:
$E\{E\{J(Y,\hat{Y}(x))|\{X_i, Y_i\}\}\}$
over all future data, given training set; over all training sets

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

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

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

- $f_k(X)$ multivariate Gaussian, same Σ per class → LDA (variance-covariance)
- $f_k(X)$ multivariate Gaussian, different Σ_k → QDA (quadratic discriminant analysis)

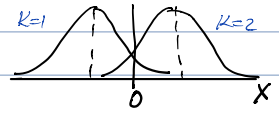
Linear discriminant analysis

$f(X)$ = 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

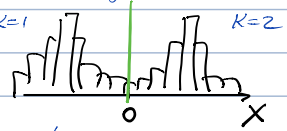
Example: Univariate Gaussian

Population: $N(\mu_k, \sigma^2)$, same



$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu_k)^2}$

Data: histograms



Q: what is the optimal decision boundary?

$P\{Y=k|X\} = \frac{\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)\cdot\pi_k}{\sum_{l=1}^{2}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)\pi_l}$

same variance; does not depend on k

Predict class k:

$$\hat{Y}(x) = \arg\max_k P\{Y=k|X\} = \arg\max_k \log P\{Y=k|X\}$$

$$= \arg\max_k \left\{ -\frac{1}{2\sigma^2}(x^2 - 2x\mu_k + \mu_k^2) + \log\pi_k \right\}$$

(x^2: does not depend on k)

$$= \arg\max_k \left\{ \frac{x\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log\pi_k \right\} = \arg\max_k \delta_k(x)$$

$\delta_k(x)$, linear in x → linear discriminant function

Assume $\pi_1 = \pi_2 = \frac{1}{2}$

$$\hat{Y}(x) = \arg\max_k \left\{ x\mu_k - \frac{1}{2}\mu_k^2 \right\}.$$

The boundary is $\{x\}$ where both classes have same posterior prob:

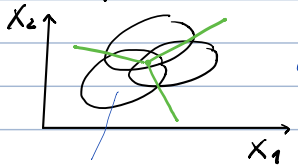
$$\delta_1(x) = x\mu_1 - \frac{1}{2}\mu_1^2 = x\mu_2 - \frac{1}{2}\mu_2^2 = \delta_2(x)$$

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}$$

In practice, estimate μ_k, σ and π_k from the data

Example: multivariate Gaussian, $K>2$ classes

Population:



curves of same shape

- multivariate Gaussian $N(\mu_k, \Sigma)$ ($p\times 1$, $p\times p$; in this case: 2×1, 2×2)

$$\hat{Y}(X)_{p\times 1} = \arg\max_k P\{Y=k|X\}$$

Bayes rule

$$= \arg\max_k \frac{f_k(X)\pi_k}{\sum_{l=1}^{K} f_l(X)\pi_l}$$

$$= \arg\max_k f_k(x)\pi_k$$

$$= \arg\max_k [\log f_k(x) + \log\pi_k]$$

Next, substitute multivariate Normal distribution:

$$\hat{Y}(X) = \arg\max_k \left\{ -\log(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}} - \frac{1}{2}\underbrace{(X-\mu_k)^T}_{1\times p}\underbrace{\Sigma^{-1}}_{p\times p}\underbrace{(X-\mu_k)}_{p\times 1} + \log\pi_k \right\}$$

($(X-\mu_k)$: $p\times 1$; 1×1)

($-\log(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}$: does not depend on k)

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

Linear expression in X → $\delta_k(X)$ - linear discriminant function

Decision boundary between classes k and l:

$$\{X: P\{Y=k|X\} = P\{Y=l|X\}\} = \{X: \delta_k(X) = \delta_l(X)\} = \{X: \delta_k(X) - \delta_l(X) = 0\}$$

$$\overset{MVN}{=} \left\{X: X^T\Sigma^{-1}(\mu_k - \mu_l) - \frac{1}{2}(\mu_k - \mu_l)^T\Sigma^{-1}(\mu_k - \mu_l) + \log\frac{\pi_k}{\pi_l} = 0\right\}$$

For example, $K=2$:

$$\log\frac{\pi_1}{\pi_2} - \frac{1}{2}(\mu_1 - \mu_2)^T\Sigma^{-1}(\mu_1 - \mu_2) \overset{\text{new notation}}{=} a_0$$

$$\underbrace{\Sigma^{-1}(\mu_1 - \mu_2)}_{p\times 1} \overset{\text{new notation}}{=} (a_1\ a_2 \dots a_p)^T$$

(Σ^{-1}: $p\times p$; $(\mu_1 - \mu_2)$: $p\times 1$)

$\Rightarrow$ classify $\hat{Y}=1$ if $a_0 + \sum_{j=1}^{p} a_j x_j > 0$

Linear combination of all predictors (i.e., no feature selection!)

Report posterior probability $P\{Y=k|X\} = \frac{f_k(x)\pi_k}{\sum_{\ell=1}^{K} f_\ell(x)\pi_\ell} = \frac{\exp\{-\frac{1}{2}\delta_k(x)\}}{\sum_{\ell=1}^{K}\exp\{-\frac{1}{2}\delta_\ell(x)\}}$

Non-linear decision boundaries: create higher order features X_1^2, X_2^2, X_1X_2 etc

Estimation:

$\hat{\pi}_k = \frac{N_k}{N}$ ← # obs in class k / ← total # obs

$\hat{\mu}_k = \begin{pmatrix} \bar{X}_{\cdot 1} \\ \bar{X}_{\cdot 2} \\ \dots \\ \bar{X}_{\cdot p} \end{pmatrix}$ ← average over all obs. of class k

$\hat{\Sigma}_{p\times p} = \frac{1}{N-K}\sum_{k=1}^{K}\sum_{\{i:\,y_i=k\}}(X_i-\hat{\mu}_k)(X_i-\hat{\mu}_k)^T$ — pooled variance-covariance across all classes

Summary of LDA:

- decision boundary is a linear combination of all features
- no feature selection
- multivariate Gaussian
- Large # of parameters (esp. when p is large and N is small)
- without assuming common Σ across classes → quadratic decision boundary (QDA)
- non-linear decision boundaries arize from (1) X^2 etc, or (2) QDA
- subject to bias-variance trade-off

LDA in high dimensions

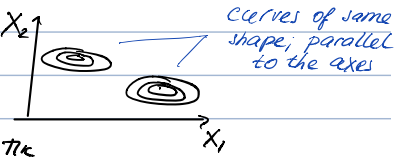
Nearest centroids (diagonal covariance LDA)

When $p \gg N$, estimation of variance-covariance is unstable

Simplification: assume that the predictors are independent

$\hat{\mu}_k = \begin{pmatrix} \bar{X}_{\cdot 1} \\ \dots \\ \bar{X}_{\cdot p} \end{pmatrix}$ over all $\{i: y_i=k\}$, $\Sigma = \begin{pmatrix} s_1^2 & & 0 \\ & s_2^2 & \\ 0 & & \ddots \\ & & & s_p^2 \end{pmatrix}$

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



The discriminant function becomes $\delta_k(x_i) = \sum_{j=1}^{p}\frac{(x_{ij}-\bar{x}_{kj})^2}{s_j^2} - 2\log\pi_k$

(x_i: new obs; $\bar{x}_{kj}$: over $\{i: y_i=k\}$; $\delta_k(x_i)$: for obs. i)

⇒ nearest centroid classifier

(after standardization and correcting for class prior)

$\hat{Y}(x_i) = \arg\max_k \delta_k(x_i)$

Regularized LDA

Middle ground between LDA and QDA

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

class-specific pooled

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)

Choose α by cross-validation

Nearest shrunken centroids:

$\hat{\mu}_k = \begin{pmatrix} \bar{x}_{\cdot 1} \\ \cdots \\ \bar{x}_{\cdot p} \end{pmatrix}$ — centroid of class k; over all $\{i : y_i = k\}$

$\hat{\mu} = \begin{pmatrix} \bar{x}_{\cdot 1} \\ \cdots \\ \bar{x}_{\cdot p} \end{pmatrix}$ — overall centroid; all i

Define $d_{jk} = \dfrac{\hat{\mu}_{jk} - \hat{\mu}_j}{(\sqrt{1/n_k - 1/N} + m_k) \cdot s_j}$

standardized distance of class centroid from overall centroid in dimension j

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

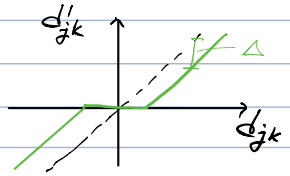
$\Rightarrow$

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

when small, dimension j is not informative for class k

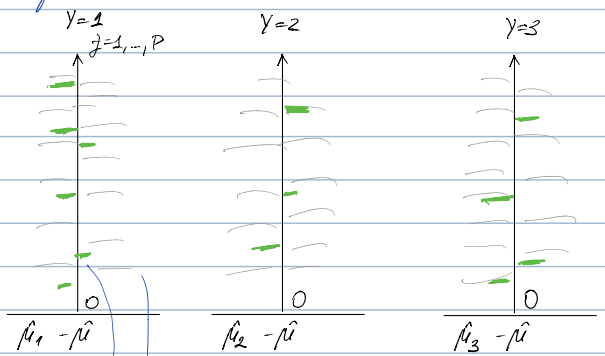
Replace with $d'_{jk} = \text{sign}(d_{jk}) \cdot \{ |d_{jk}| - \Delta \}_+$ (choose by cross-validation)

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



soft threshold function

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



original centroids

shrunken centroids (smaller; many = 0)

Then the discriminant function is

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

x_j: new data point; $\hat{\mu}'_{jk}$: shrunken centroid

Class probabilities:

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