• Probability

- Bernoulli (π) $\quad E=\pi$

$P(Y=y)=\begin{cases}\pi & y=1\\ 1-\pi & y=0\end{cases}$ $\quad Var=\pi(1-\pi)$

- Uniform $U(a,b)$ $\quad E=\frac{a+b}{2}$ $\quad Var=\frac{1}{12}(b-a)^2$

$P(Y=y)=\begin{cases}\frac{1}{b-a} & \text{if } y\in[a,b]\\ 0\end{cases}$

- Binomial (n,π) $\quad E=n\pi,\ Var=n\pi(1-\pi)$

$P(Y=y)=\binom{n}{y}\pi^y(1-\pi)^{n-y}$

- normal distribution $N(\mu,\sigma)$ $\quad E=\mu$ $\quad Var=\sigma^2$

$P(Y=y)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{Y-\mu}{\sigma}\right)^2}$

- Poisson (λ) $\quad E=\lambda$ $\quad Var=\lambda$

$P(Y=y)=\frac{e^{-\lambda}\lambda^y}{y!}$

- standard normal distribution $N(0,1)$

$P(Y=y)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}$

• conditional ~~prop~~ probability

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

- Bayes rule

$P(Y|X)=\frac{P(X|Y)\times P(Y)}{P(X)}=\frac{P(X|Y)\times P(Y)}{\Sigma_y P(y)\cdot P(X|y)}$

example.

$P(HIV|pos)=\frac{P(pos|HIV)\times P(HIV)}{P(pos)}=\frac{P(pos|HIV)\times P(HIV)}{P(pos|HIV)\times P(HIV)+P(pos|\text{not }HIV)\times P(\text{not }HIV)}$

- independence

two events are independent if $P(Y|X)=P(Y)$

$P(X,Y)=P(X)\times P(Y|X)=P(X)\times P(Y)$

- pdf vs cdf probability density function cumulative distribution function

cdf $F_Y(y)=P(Y\le y)$ ordered.

$=\int_{-\inf}^{y}f_Y(y)dx$ continuous

- For random variable $Y\sim N(\mu,\sigma)$

$P(Y\le y)=P\left\{\frac{Y-\mu}{\sigma}\le\frac{y-\mu}{\sigma}\right\}=P\left\{z\le\frac{y-\mu}{\sigma}\right\}$ 查表

- Bivariate Normal distribution (Y_1,Y_2)

requires five parameters

$\mu_1,\sigma_1,\mu_2,\sigma_2,\rho_{12}$

$\rho_{12}=\frac{cov(Y_1,Y_2)}{\sigma_1\sigma_2}=\frac{E[(Y_1-\mu_1)(Y_2-\mu_2)]}{\sigma_1\sigma_2}=\frac{E(Y_1Y_2)-E(Y_1)E(Y_2)}{\sigma_1\sigma_2}$

↑

coefficient of correlation

- ★ Uncorrelated ≠ independent

independent → uncorrelated

uncorrelated ↛ independent.

- $E(Y_1|Y_2)=(\mu_1-\mu_2\rho_{12}\frac{\sigma_1}{\sigma_2})+\rho_{12}\frac{\sigma_1}{\sigma_2}Y_2=\alpha_{1|2}+\beta_{1|2}Y_2$

$Var(Y_1|Y_2)=\sigma_1^2(1-\rho_{12}^2)$

• Information theory

- Entropy → measure of its uncertainty

$H(y)=-\sum_{c=1}^{C}p(y=c)\log_2 p(y=c)$

- Kullback-Leibler divergence

- dissimilarity of two prob. distributions p and q

$KL(p,q)=\sum_{c=1}^{C}p_c\cdot\log_2\frac{p_c}{q_c}=\sum_{c=1}^{C}p_c\log_2 p_c-\sum_{c=1}^{C}p_c\log_2 q_c=-H(p)+H(p,q)$

$=-\text{entropy}+\text{cross-entropy}$

- Mutual information

- MI = 0 iff the variables are independent

$I(x,y)=\sum_x\sum_y p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)}$

• Sampling distribution

- Variability of summaries of random samples from the population (mean, variance)

1. $y_1,y_2\ldots y_n$ collect from population
2. calculate mean $\bar{y}$
3. repeat 1, 2 large times
4. The histogram of large value of $\bar{y}$ approximates the sampling distribution of $\bar{Y}$.

• CLT central

- $y_1,y_2,\ldots,y_n$ follow arbitrary probability distribution with expected value μ and sd σ and n is large.

- then $\hat{Y}\sim N(\mu,\frac{\sigma}{\sqrt{n}})$

★★ difference between standard deviation and standard error.

sd: quantifies the variability in the population

SE: - - - - - - uncertainty in a parameter of population

sd remains the same when sample size ↑

whe SE ↓

$SE=\frac{sd}{\sqrt{n}}$

Linear regression (discriminative classifiers) $\quad Y_i\sim N(\beta_0+\beta_1X_{i1}+\beta_2X_{i2},\sigma^2)$

$Y_i=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\varepsilon_i,\ i=1,\ldots,n\quad \varepsilon_i\overset{iid}{\sim}N(0,\sigma^2)$ $\quad Y=X\beta+\varepsilon$ matrix form

- β_1 describes change in mean response per unit increase in X_1 when X_2 is held constant
- β_2 describes change in mean response per unit increase in X_2 when X_1 is held constant

• Interaction model

$Y_i=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\beta_3X_{i1}X_{i2}+\varepsilon_i,\ i=1,\ldots,n\quad \varepsilon_i\overset{iid}{\sim}N(0,\sigma^2)$

- change in X_1 when $X_2=x_2$ $\quad \Delta Y=\beta_1+\beta_3x_2$
- change in X_2 when $X_1=x_1$ $\quad \Delta Y=\beta_2+\beta_3x_1$

Rate of change due to one variable affected by the other.

• Least squares Estimation same as ~~or~~ maximum likelihood estimation

- minimize $(Y-X\hat{\beta})^T(Y-X\beta)$

$\hat{\beta}=(X^TX)^{-1}X^TY$

$\therefore\hat{Y}=X\hat{\beta}=\{X(X^TX)^{-1}X^T\}Y=HY$

- residuals

$e=Y-\hat{Y}=(I-H)Y$

- $f(i)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(Y_i-\beta_0-\beta_1X_i)^2}$

$L(\theta)=f_1\times f_2\times\ldots\times f_n=\sum_{i=1}^{n}\log\left[\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{1}{2}}e^{-\frac{1}{2\sigma^2}(y_i-\beta_0-\beta_1X_i)^2}\right]=-\frac{1}{2\sigma^2}RSS(\theta)-\frac{n}{2}\log(2\pi\sigma^2)$

- Find $\theta(\beta_0,\beta_1\ldots\beta_p)$ and σ^2 which maximizes L = minimize $(Y-X\hat{\beta})^T(Y-X\hat{\beta})$

• Qualitative predictors

- $Y_i=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\beta_3X_{i1}X_{i2}+\varepsilon_i$

$X_2=1$ if case from MA

- Three groups

$Y_i=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+\beta_3X_{i3}+\varepsilon_i$

$X_2=1$ if stock firm

$X_3=1$ if foreign firm

with X_1X_2 and X_1X_3 terms

more flexible, interaction model.

• Partition sums of squares

- Total sums of squares

$SSTO=\Sigma(Y_i-\bar{Y})^2=\Sigma(Y_i-\hat{Y}_i+\hat{Y}_i-\bar{Y})^2=\Sigma(\hat{Y}_i-\bar{Y})^2+\Sigma(Y_i-\hat{Y}_i)^2$

$=b^2\Sigma(X_i-\bar{X})^2+\Sigma(Y_i-\hat{Y}_i)^2$

- $SSTO=\underset{\text{regression sum of squares}}{SSR}+\underset{\text{RSS residual sum of squares}}{SSE}$

R^2 is not that usefull

• Accessing the accuracy of the model

$R^2=\frac{SSR}{SSTO}=1-\frac{SSE}{SSTO}$ ← measure the proportion of variability than can be explained using X. Adjust $R^2=1-\frac{SSE/n-p}{SSTO/n-1}$ wants to account for p

- The number of indicators is always one less than the number of groups.

• Gauss-Markov theorem

Least squares estimator has the smallest mean squared error of all linear estimators with no bias.

• Model selection and bias-variance trade off

- Loss function $L(Y,f(X))$
- risk $R=E[L(Y,f(X))]=\sum_x\sum_y L(Y,f(X))\cdot p(x,y)$
- estimated risk $R=\frac{1}{N}\sum_{i=1}^{N}L(Y,f(X))$

➤ $Y=f_k(X)+\varepsilon,\ L=(Y-f_k(X))^2$

$E[(Y-f_k(X)|X=x)^2]=\sigma^2+\left[f(x)-\frac{1}{n'}\sum_{i=1}^{n'}f_k(x_i)\right]^2+\frac{\sigma^2}{n'}$

$=\text{irreducible error}+Bias^2(\hat{f}_k(x))+Var(\hat{f}_k(x))$

• Bias-variance decomposition

$B(\hat{\theta})=E(\hat{\theta}-\theta)=E(\hat{\theta})-\theta$ Bias

$Var(\hat{\theta})=E[(\hat{\theta}-E(\hat{\theta}))^2]$ variance.

- Mean squared Error (MSE)

$MSE(\hat{\theta})=E[(\hat{\theta}-\theta)^2]$

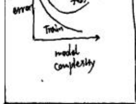
$=E[(\hat{\theta}-E(\hat{\theta})+E(\hat{\theta})-\theta)^2]$

$=E[(\hat{\theta}-E(\hat{\theta}))^2+(E(\hat{\theta})-\theta)^2+2(\hat{\theta}-E(\hat{\theta}))(E(\hat{\theta})-\theta)]$

$\because E(\hat{\theta}-E(\hat{\theta}))=E(\hat{\theta})-E(\hat{\theta})=0$

又$\because E(\hat{\theta})-\theta$ is constant $\quad$ we drop the cross term

$MSE(\hat{\theta})=Var(\hat{\theta})+B(\hat{\theta})^2$



• AIC and BIC criterions to chose between models

$-2\log L+kp$

$AIC=\frac{SSE}{MSE}+2p$ or $n\log(\frac{SSE}{n})+2p$

$BIC=\frac{SSE}{MSE}+p\log(n)$ (heavier penalty) $\quad n\log(\frac{SSE}{n})+\log(n)\cdot p$

• cross-validation

Iteratively use each part for training/ variable selection/ validation

★ Cross-validation should be done as part of variable selection.

- Ridge Regression — scaled version of least squares $\hat{\beta}/(1+\lambda)$
- motivation
  difficulty inverting $X'X$ when multicollinearity
- minimize
  $RSS + \lambda \sum_{j=1}^{p} \beta_j^2$ — use CV to choose λ — $(Y^* - X^*\beta^*)'(Y^* - X^*\beta^*) + \lambda \cdot \beta^{*\prime}\beta^*$
- $\hat{\beta} = (X^{*\prime}X^* + \lambda I)^{-1} X^{*\prime} Y^*$
  biased but stable

elastic net penalty
$\lambda \sum_{j=1}^{p} [\alpha \beta_j^2 + (1-\alpha)|\beta_j|]$

- Lasso Regression — soft thresholding $\text{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda)_+$
- minimize
  $RSS + \lambda \sum_{j=1}^{p} |\beta_j|$ — weaker penalty on β than Ridge.
  or
  minimize
  $\sum_{i=1}^{n} [Y_i^* - (\beta_0^* + \beta_1^* X_{i1}^* + \dots + \beta_p^* X_{ip}^*)]^2$ s.t. $\sum_{j=1}^{p} |\beta_j| \le t$
  optimum is on the intersection with some axes

$\lambda \to \infty$
training residual sum of squares ↑
Test residual ↓ then ↑
variance ↓
squared bias ↑

- Logistic regression — discriminative classifiers
- $E(Y_i) = g(\beta_0 + \beta_1 X_i)$ or $g^{-1}(E(Y_i)) = \beta_0 + \beta_1 X_i$
  g: mean response function g^{-1}: link function
  $g(t)$: identity → linear regression — w/ does not meet $0 \le E(Y_i) \le 1$; non-normal distribution of ε; non-constant variance
  $\Phi(t)$ → probit
  $\frac{e^{(t)}}{1+e^{(t)}}$ → logistic
- motivation for probit regression: Latent variable
  $bp = \beta_0 + \beta_1 age + \varepsilon$
  $P(Y_i = 1) = P(bp > c) = P(\beta_0 + \beta_1 age + \varepsilon > c) = P(\varepsilon < \beta_0 + \beta_1 age - c)$
  $= P(\frac{\varepsilon}{\sigma} < \frac{\beta_0 - c}{\sigma} + \frac{\beta_1}{\sigma} age) = P(z < \beta_0' + \beta_1' age) = \Phi(\beta_0' + \beta_1' age)$
- sigmoidal response function
  $E(Y_i) = \frac{e^{\beta_0 + \beta_1 X_i}}{1 + e^{\beta_0 + \beta_1 X_i}}$
  Logit link function — odds
  $\log\left(\frac{E(Y_i)}{1 - E(Y_i)}\right) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 X_i$

- confusion matrix

		predicted class −	+			
True	−	TrueNeg	FalsePos (Type I error)	N	$\frac{TP}{P}$	sensitivity
class	+	FalseNeg (Type II error)	TruePos	P	$\frac{FP}{N}$	1-specificity
		N^*	P^*			

- ROC curve Receiver Operating Characteristic
  simultaneously displaying the two types of errors for all possible thresholds
  Y: true pos sensitivity — X: false pos 1-specificity
- $AIC = -2\log L(\hat{\theta}) + 2p$ — $BIC = -2\log L(\hat{\theta}) + p\log(n)$
- statistical regularization
  For logistic regression, maximize a penalized version
  Lasso $\max_{\beta_0, \beta} \left\{ \sum_{i=1}^{N} [y_i(\beta_0 + \beta^T x_i) - \log(1 + e^{\beta_0 + \beta^T x_i})] - \lambda \sum_{j=1}^{p} |\beta_j| \right\}$

- Nearest centroids
  (diagonal-covariance LDA)
  For $p >> N$, covariance estimation is unstable
  assume independent features (diagonal Σ)
- $\delta_k(X_i) = -\frac{1}{2}(X - \mu_k)' \Sigma^{-1} (X - \mu_k) + \log \pi_k$
  ⇓
  $\delta_k(X_i) = -\sum_{j=1}^{p} \frac{(X_{ij} - \bar{X}_{jk})^2}{S_j^2} + 2\log \pi_k$ — standardization; correcting class prior probability
  for class k: $\bar{X}_{jk} = \sum_{i: i \in C_k} X_{ij} / N_k$
  $\bar{X}_j = \sum_{i=1}^{N} X_{ij} / N$
  observations $i = 1, \dots, N$
  features $j = 1, \dots, p$
  $\hat{Y}(X_i) = \arg\max_k \delta_k(X_i)$
- The diagonal-covariance LDA classifier is equivalent to nearest centroid classifier after standardization, correcting for class prior probability

SVD singular value decomposition
$X = U \Lambda V'$
$n \times p$ — $n \times p$ $p \times p$ $p \times p$
↓
$XV = U\Lambda$
- Scores
  The position of each observation in the new coordinate system of principal components
  $\sum_{k=1}^{p} (X_{ik} - \bar{X}_{\cdot k}) \cdot V_{kj}$ → coordinate of observation i in direction j
- Loadings
  weight V_{kj} of feature k in the direction j
  Scores and Loadings cannot be used as "measures of statistical significance.

- PCA of X — SVD on covariance matrix is PCA (when $p >> N$)
  $\hat{\sigma}_{jk}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_{ij} - \bar{X}_{\cdot j})(X_{ik} - \bar{X}_{\cdot k})$
  $(N-1)\Sigma = X'X = V\Lambda U' U \Lambda V' = V\Lambda^2 V'$
- eigenvalues of covariance matrix $= \frac{1}{N-1} \lambda_j^2$

- MLE — independent but not identically distributed.
  $Y_i \overset{ind}{\sim} \text{Bernoulli}(\pi_i)$ — $\pi_i = \frac{e^{(t)}}{1+e^{(t)}}$
  $f(Y_i) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$
  $\log L = \log \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$
  $= \sum_{i=1}^{n} Y_i \log(\pi_i) + \sum_{i=1}^{n} (1-Y_i)\log(1-\pi_i)$
  $= \sum_{i=1}^{n} Y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \sum_{i=1}^{n} \log(1-\pi_i)$
  $= \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i) - \sum_{i=1}^{n} \log(1 + e^{\beta_0 + \beta_1 X_i})$

  $Y_i' \overset{ind}{\sim} \text{Binomial}(n_i, \pi_i)$
  $\log L = \log \prod_{i=1}^{c} \left(\binom{n_i}{Y_i'} \pi_i^{Y_i'} (1-\pi_i)^{n_i - Y_i'} \right)$
  $= \sum_{i=1}^{c} [Y_i' \log(\pi_i) + (n_i - Y_i')\log(1-\pi_i) + \log\binom{n_i}{Y_i'}]$
  $= \sum_{i=1}^{c} [Y_i' \log\frac{\pi_i}{1-\pi_i} + n_i \log(1-\pi_i) + \log\binom{n_i}{Y_i'}]$
  $= \log L^{bernoulli} + \text{constant}$
  $L^{bernoulli}$ and $L^{binomial}$ lead same parameter estimates and inferences but different deviances
  $D(y) = -2(\log(p(y|\hat{\theta}_0)) - \log(p(y|\hat{\theta}_s)))$

★ Variable selection should be done as part of cross-validation!

$Y | X = x \sim \text{Multinomial}(n_x, \pi_1(x), \dots, \pi_J(x))$
$\log \frac{\pi_j(x)}{\pi_1(x)} = \alpha_j + \beta_j x$
$\pi_j(x) = \pi_1(x) e^{\alpha_j + \beta_j x}$ — $\sum_{j=1}^{J} \pi_j(x) = 1$
$\pi_1(x) = \frac{1}{1 + \sum_{k=2}^{J} e^{\alpha_k + \beta_k x}}, j = 1$
$\pi_j(x) = \frac{e^{\alpha_j + \beta_j x}}{1 + \sum_{k=2}^{J} e^{\alpha_k + \beta_k x}}, j = 2, 3, \dots, J$
For $J = 2$, → ordinary logistic regression

Data:
(y_{ij}, x_i);
$y_{ij} = I(y_i = j)$
$\sum_{j=1}^{J} y_{ij} = 1$
$y_{i1} = (1 - \sum_{j=2}^{J} y_{ij})$

- Comments on PCA
- Decomposition of Σ helps interpretation
  $\sum_{j=1}^{p}$ sample variance of columns of X:
  $\sum_{j=1}^{p} \frac{1}{N-1} \sum_{i=1}^{N} (X_{ij} - \bar{X}_{\cdot j})^2 = \frac{1}{N-1} tr(X'X)$
  $= \frac{1}{N-1} tr(V\Lambda^2 V') = \frac{1}{N-1} tr(\Lambda^2 V'V)$
  $= \frac{1}{N-1} tr \Lambda^2 = \frac{\sum_{j=1}^{p} \lambda_j^2}{N-1}$
- The proportion of total variance explained by the principle component j is $\frac{\lambda_j^2}{\sum \lambda_k^2}$
- doesn't distinguish useful and nuisance variation — between-group, within-group
- can be driven by large nuisance variation
- can not use scores as evidence of the ability of the features
- PCA performs best when all features have similar nuisance variation

- Inversion of Σ^{-1} in $\delta_k(X)$
  For LDA discriminant function is
  $\delta_k(X) = -\frac{1}{2}\log|\Sigma| - \frac{1}{2}(X - \mu_k)' \Sigma^{-1} (X - \mu_k) + \log \pi_k$
  $\Sigma^{-1} = (V\Lambda^2 V')^{-1}$
  $\therefore (X - \mu_k)' \Sigma^{-1} (X - \mu_k) = [\Lambda^{-1} V' (X - \mu_k)]' [\Lambda^{-1} V' (X - \mu_k)]$
  ① sphere the data
  $\Lambda^{-1} V' X \to X^*$ — $\Lambda^{-1} V' \mu_k \to \mu^*$
  ② classify to closest class centroid.

- Fisher's approach
  project high-dimensional data onto a lower-dimensional space to best separate class
  $\max \frac{\text{between-group variance}}{\text{within-group variance}} = \max_a \frac{a'Ba}{a'\Sigma a}$
- transform the system coordinates s.t. $a'\Sigma a = 1$
- maximize $a'Ba$ in the transformed system

- Regularized LDA regularization
- $\hat{\Sigma}_k(\alpha) = \alpha \hat{\Sigma}_k + (1-\alpha)\hat{\Sigma}$ shrink the separate covariances of QDA toward a common covariance as in LDA
- Shrink the common covariance in LDA toward a scalar covariance
  $\hat{\Sigma}(\alpha) = \alpha \hat{\Sigma} + (1-\alpha)\sigma^2 I$

- MLE
  $l_i = \log \prod_{j=1}^{J} \pi_j(x_i)^{y_{ij}}$
  $= \sum_{j=2}^{J} y_{ij} \log \pi_j(x_i) + (1 - \sum_{j=2}^{J} y_{ij}) \log \pi_1(x_i)$
  $= \sum_{j=2}^{J} y_{ij} \log \frac{\pi_j(x_i)}{1 - \sum_{k=2}^{J} \pi_k(x_i)} + \log \pi_1(x_i)$
  $= \sum_{j=2}^{J} y_{ij}(\alpha_j + \beta_j x_i) - \log[1 + \sum_{j=2}^{J} e^{\alpha_j + \beta_j x_i}]$
  maximize $\sum_{i=1}^{n} l_i$ with respect to α_j and β_j
- can fit $J-1$ logistic regressions for $J-1$ response categories vs baseline (the same model)

- Generative vs discriminative
  $P(Y|X) = \frac{P(Y) \cdot P(X|Y)}{P(X)}$
  generative classifiers
  - specify prior probability $P(Y)$
  - Assume conditional distribution $P(X|Y)$
  - use Bayes rule to derive the posterior $P(Y|X)$
  example: LDA
  discriminative
  - estimate posterior $P(Y|X)$ directly
  example: linear regression, logistic regression

- connection of LDA and logistic regression
- Linear decision boundaries
- logistic regression estimated using MLE
- coefficients in LDA are estimated using $\hat{\mu}$ and $\hat{\sigma}$ from MVN

$P(Y = k | X) = \frac{f_k(X)\pi_k}{\sum_{l=1}^{K} f_l(X)\pi_l}$
MAP decision
$\hat{Y}(X) = \arg\max_k P(Y = k | X) = \arg\max_k f_k(X)\pi_k$
$f_k(X)$ types:
- Gaussian, same Σ → LDA
- Gaussian, different Σ_k → QDA
- mixture of Gaussians → Mixture models
- $\prod_{p=1}^{P} f_k(X_p)$ every variable independent Naive Bayes
- $f_k(X)$ nonparametric → Kernel estimates.

- Nearest regular centroid regularization
  $d_{ik} = \frac{\bar{X}_{ik} - \bar{X}_i}{m_k \cdot S_i}$ for feature i, class k
  $S_i^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{j \in C_k} (X_{ij} - \bar{X}_{ik})^2$
  $m_k = \sqrt{1/n_k - 1/n}$
- Nearest shrunken centroids
  Start with nearest centroids
  $\bar{X}_{ik} = \bar{X}_i + m_k \cdot S_i \cdot d_{ik}$
  ↓
  $\bar{X}_{ik}' = \bar{X} + m_k \cdot S_i \cdot d_{ik}'$ shrink d_{ik} to 0
  $d_{ik}' = \text{sign}(d_{ik})(|d_{ik}| - \Delta)_+$
  Δ is chosen by CV (log-likelihood or error rate)

- Soft thresholding vs hard
  $d_{ik}' = d_{ik} \cdot I(|d_{ik}| > \Delta)$
  we prefer soft-thresholding as it is a smoother operation and works better and not jumpy.
  discriminant score
  $\delta_k(X^*) = \sum_{i=1}^{p} \frac{(X_i^* - \bar{X}_{ik}')^2}{S_i^2} - 2\log \pi_k$
  ↑
  standardized squared distance of X^* to the kth shrunken centroid.

- LDA one predictor
  $f_k(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(X - \mu_k)^2\right)$
  $\hat{Y}(X) = \arg\max_k P_r(Y = k | X)$
  $= \arg\max_k \left[X \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)\right]$
  $= \arg\max_k \delta_k(X)$
- MVN distribution & Σ
  $\hat{Y}(X) = \arg\max_k P_r(Y = k | X)$
  $= \arg\max_k \frac{f_k(X)\pi_k}{\sum_{l=1}^{K} f_l(X)\pi_l}$
  $= \arg\max_k f_k(X)\pi_k$
  $= \arg\max_k [\log f_k(X) + \log \pi_k]$
  $= \arg\max_k [-\log(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}} - \frac{1}{2}(X - \mu_k)' \Sigma^{-1} (X - \mu_k) + \log \pi_k]$
  $= \arg\max_k [-\frac{1}{2}(X - \mu_k)' \Sigma^{-1} (X - \mu_k) + \log \pi_k]$
  $= \arg\max_k [X' \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k' \Sigma^{-1} \mu_k + \log \pi_k]$
- decision boundary for LDA
  $P_r\{Y = k | X\} = P_r\{Y = l | X\}$
  $\Rightarrow X: \log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k + \mu_l)' \Sigma^{-1} (\mu_k - \mu_l) + X' \Sigma^{-1} (\mu_k - \mu_l) = 0$
- QDA or LQA Quadratic discriminant Analysis
  $\hat{Y}(X) = \arg\max_k f_k(X)\pi_k = \arg\max_k [-\log(2\pi)^{\frac{p}{2}} |\Sigma_k|^{\frac{1}{2}} - \frac{1}{2}(X - \mu_k)' \Sigma_k^{-1} (X - \mu_k) + \log \pi_k]$
- decision boundary
  $X: \log \frac{\pi_k}{\pi_l} - \log \frac{|\Sigma_k|}{|\Sigma_l|} - \frac{1}{2}(X - \mu_k)' \Sigma_k^{-1} (X - \mu_k) + \frac{1}{2}(X - \mu_l)' \Sigma_l^{-1} (X - \mu_l) = 0$
- Estimate the decision boundary
  $\hat{\pi}_k = N_k / N$
  $\hat{\mu}_k = \sum_{i: Y_i = k} X_i / N_k$
  $\hat{\Sigma} = \sum_{k=1}^{K} \sum_{i: Y_i = k} (X_i - \hat{\mu}_k)(X_i - \hat{\mu}_k)' / (N - K)$
- $(K-1)(p+1)$ parameters for LDA
  $(K-1)(p(p+3)/2 + 1)$ parameters for QDA

★ Likelihood vs probability
- Probability is used before data are available to describe future outcomes given a fixed value for the parameter.
- Likelihood is used after data are available to describe a function of a parameter for a given outcome.

★ Why not least squares in logistic regression?
In logistic regression, the errors are not expected to have the same variance: should have high var when p near .5, low variance towards extremes. Leads to IRWLS but not simple LS. to
Iteratively weighted least squares