Generative modeling

Hastie, Tibshirani, Friedman Ch 4.3

Kevin Murphy Ch. 4.1-4.2

CS 6140
Machine Learning
Professor Olga Vitek

February 16, 2017

Generative vs discriminative models

- Goal: predict Y
  - Bayes rule:

$$p(Y|\mathbf{X}) = \frac{p(Y) \cdot p(\mathbf{X}|Y)}{p(\mathbf{X})}$$

- Generative classifiers
  - Specify prior probability of $p(Y)$
  - Assume conditional distribution $p(\mathbf{X}|Y)$
  - Use Bayes rule to derive the posterior $p(Y|\mathbf{X})$
  - **Example:** Linear discriminant analysis

- Discriminative classifiers
  - Estimate the posterior the posterior $p(Y|\mathbf{X})$
  - Do not assume the distribution on $\mathbf{X}$
  - **Example:** Y binary: logistic regression

Generative modeling

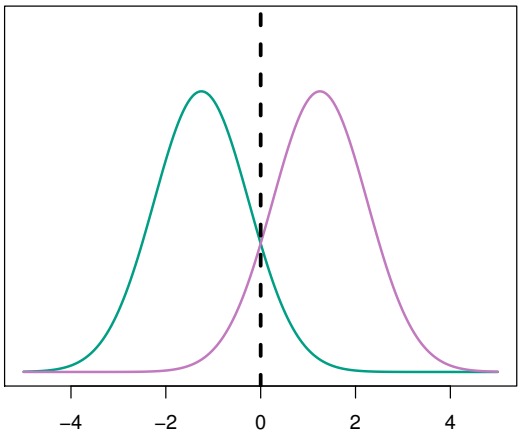
- Priors π_k, $\sum_k = 1$.
  - Usually $\hat{\pi}_k = \frac{\text{\# of observations in class } k}{\text{Total \# of observations}}$
- Density $f_k(\mathbf{X})$
- Posterior $\Pr\{Y = k|\mathbf{X}\} = \frac{f_k(\mathbf{X})\pi_k}{\sum_{l=1}^{K} f_l(\mathbf{X})\pi_l}$
- MAP (maximum *a posteriori*) decision
  - Based on Bayes rule under 0-1 loss

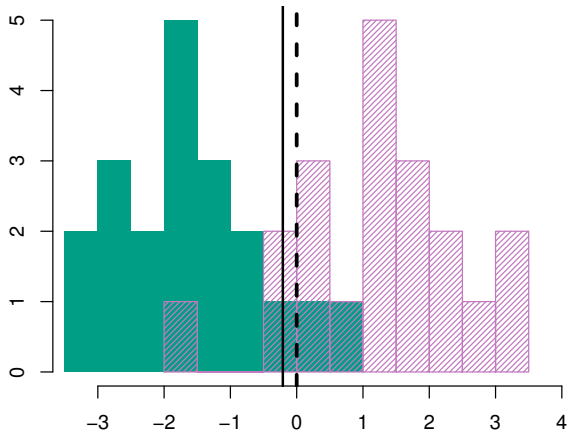
    $\hat{Y}(\mathbf{X}) = \arg\max_k \Pr\{Y = k|\mathbf{X}\} = \arg\max_k f_k(\mathbf{X})\,\pi_k$

- Specification of $f_k(\mathbf{X})$ defines the classifier
  - $f_k(\mathbf{X})$ Gaussian, same Σ per class $\longrightarrow$ LDA
  - $f_k(\mathbf{X})$ Gaussian, different $\Sigma_k \longrightarrow$ LQA
  - $f_k(\mathbf{X})$ mixture of Gaussians $\longrightarrow$ Mixture models
  - $f_k(\mathbf{X}) = \prod_{p=1}^{P} f_k(X_p) \longrightarrow$ Naïve Bayes
  - $f_k(\mathbf{X})$ nonparametric $\longrightarrow$ Kernel estimates

Linear Discriminant Analysis

One predictor: model





Assume that the observations are from univariate Normal distributions with same variance

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

James, Witten, Hastie, Tibshirani
An Introduction to Statistical Learning 2013, Sec 4.4

One predictor: decision

- Plug in Normal distribution in $\Pr\{Y = k|x\}$

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

- Assign Y to class k:

$$\begin{aligned}
\hat{Y}(\mathbf{x}) &= \arg\max_k \Pr\{Y = k|x\} \\
&= \arg\max_k log(\Pr\{Y = k|x\}) \\
&= \arg\max_k x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + log(\pi_k) \\
&= \arg\max_k \delta_k(x)
\end{aligned}$$

- $\delta_k(x)$ is linear in x

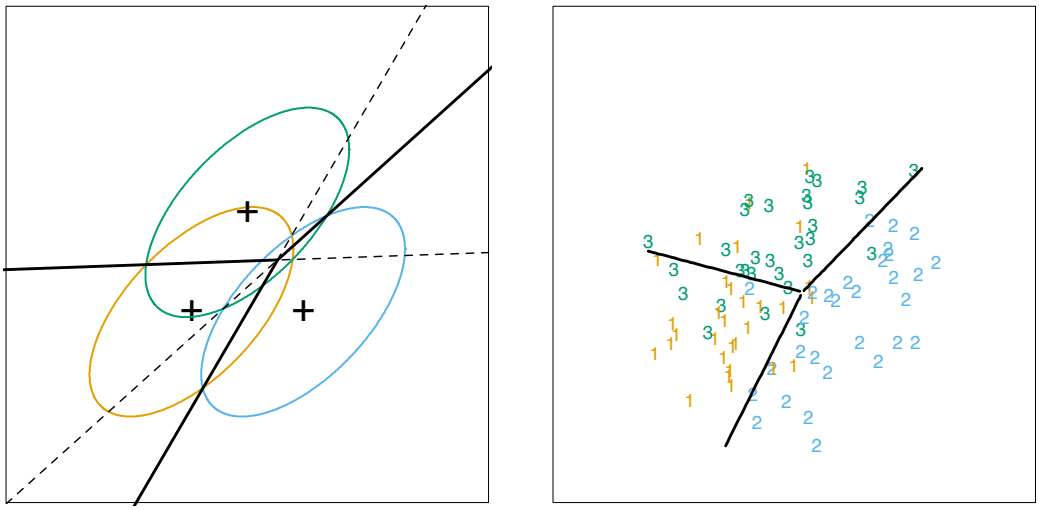
- When $\pi_1 = \pi_2$, the decision boundary is

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

- In practice, estimate π_k, μ_k, σ from data

James, Witten, Hastie, Tibshirani
An Introduction to Statistical Learning 2013, Sec 4.4

Two predictors



Assume that the observations are from a multivariate Normal distributions with same variance-covariance

- Left: (P=2)
  - Contours: 95% of the probability.
  - Dashed lines: true Bayes decision boundaries between pairs of classes
  - Solid lines: true Bayes decision boundaries between all three classes.

- Right: (P=2)
  - 20 observations from each Gaussian distribution, and the fitted LDA decision boundaries.

James, Witten, Hastie, Tibshirani
An Introduction to Statistical Learning 2013, Sec 4.4

Optimal classification

MVN distribution & equal covariance across classes

$\rightarrow$ linear rule: $\widehat{Y}(\mathbf{X}) = \arg\max_k \Pr\{Y = k|\mathbf{X}\} =$

$$\begin{aligned}
&\overset{\text{Bayes}}{=} && \arg\max_k \frac{f_k(\mathbf{X})\,\pi_k}{\sum_{l=1}^{K} f_l(\mathbf{X})\,\pi_l} \\
&\overset{\text{no den.}}{=} && \arg\max_k f_k(\mathbf{X})\,\pi_k \\
&\overset{\log}{=} && \arg\max_k log\left[f_k(\mathbf{X})\,\pi_k\right] \\
&\overset{\text{sum}}{=} && \arg\max_k \left[log f_k(\mathbf{X}) + log\pi_k\right] \\
&\overset{\text{MVN}}{=} && \arg\max_k \left[-log(2\pi)^{\frac{P}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}\right. \\
& && \left. -\frac{1}{2}(\mathbf{X}-\mu_k)'\mathbf{\Sigma}^{-1}(\mathbf{X}-\mu_k) + log\pi_k\right] \\
&\overset{\text{no const.}}{=} && \arg\max_k \left[-\frac{1}{2}(\mathbf{X}-\mu_k)'\mathbf{\Sigma}^{-1}(\mathbf{X}-\mu_k) + log\pi_k\right] \\
&\overset{\text{open}}{=} && \arg\max_k \left[\mathbf{X}'\mathbf{\Sigma}^{-1}\mu_k - \frac{1}{2}\mu_k\mathbf{\Sigma}^{-1}\mu_k - \frac{1}{2}\mathbf{X}'\mathbf{\Sigma}^{-1}\mathbf{X} + log\pi_k\right] \\
&\overset{\text{with } k}{=} && \arg\max_k \left[\mathbf{X}'\mathbf{\Sigma}^{-1}\mu_k - \frac{1}{2}\mu_k\mathbf{\Sigma}^{-1}\mu_k + log\pi_k\right]
\end{aligned}$$

The expression in brackets is linear in $\mathbf{X}$

Decision boundary

- Linear discriminant function

$$\delta_k(\mathbf{X}) = \mathbf{X}'\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k\Sigma^{-1}\mu_k + log\pi_k$$

- Decision boundary between classes k and l
  - $\{\mathbf{X}:\ \Pr\{Y=k|\mathbf{X}\} = \Pr\{Y=l|\mathbf{X}\}\}$

$$\begin{aligned} &= \{\mathbf{X}:\ \delta_k(\mathbf{X}) = \delta_l(\mathbf{X})\} \\ &= \{\mathbf{X}:\ log\delta_k(\mathbf{X}) - log\delta_l(\mathbf{X}) = 0\} \\ &= \Big\{\mathbf{X}:\ log\frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k+\mu_l)'\Sigma^{-1}(\mu_k-\mu_l) \\ &\qquad + \mathbf{X}'\Sigma^{-1}(\mu_k-\mu_l) = 0\Big\} \end{aligned}$$

  - E.g., for $K=2$

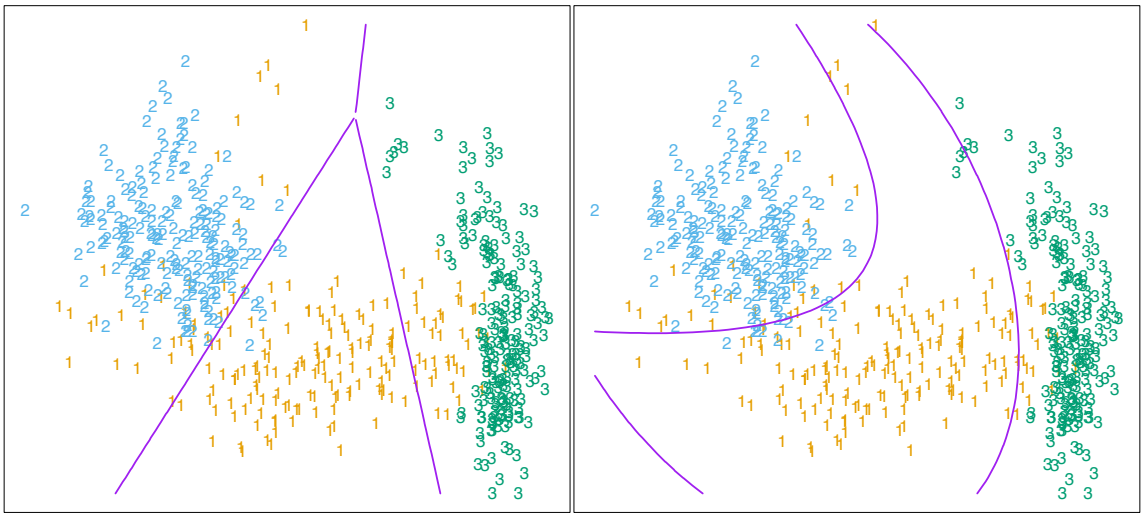
$$\begin{aligned} a_0 &= log\frac{\pi_1}{\pi_2} - \frac{1}{2}(\mu_1+\mu_2)'\Sigma^{-1}(\mu_1-\mu_2) \\ (a_1, a_2, \ldots, a_P)' &= \Sigma^{-1}(\mu_1-\mu_2) \end{aligned}$$

    Classify to class 1 if $a_0 + \sum_{j=1}^{P} a_j x_j > 0$

- Posterior probability

$$\Pr\{Y=k|\mathbf{X}\} = \frac{f_k(\mathbf{X})\,\pi_k}{\sum_{l=1}^{K} f_l(\mathbf{X})\,\pi_l} = \frac{\exp\{-\frac{1}{2}\,\delta_k(\mathbf{X})\}}{\sum_{l=1}^{K}\exp\{-\frac{1}{2}\,\delta_l(\mathbf{X})\}}$$

Non-linear decision boundaries with LDA



Can obtain non-linear decision boundaries with LDA in an augmented space

- Left: (P=2)
  - Boundaries by linear discriminant analysis
- Right: (P=2)
  - Linear boundaries in five-dimensional space $X_1, X_2, X_1X_2, X_1^2, X_2^2$
  - Linear inequalities in the augmented space are quadratic inequalities in the original space.

Hastie, Tibshirani, Friedman
The Elements of Statistical Learning 2008

Quadratic Discriminant Analysis

- Σ_k are not assumed equal
- The discriminant function is not linear

$$\hat{Y}(\mathbf{X}) = \arg\max_k f_k(\mathbf{X})\,\pi_k =$$

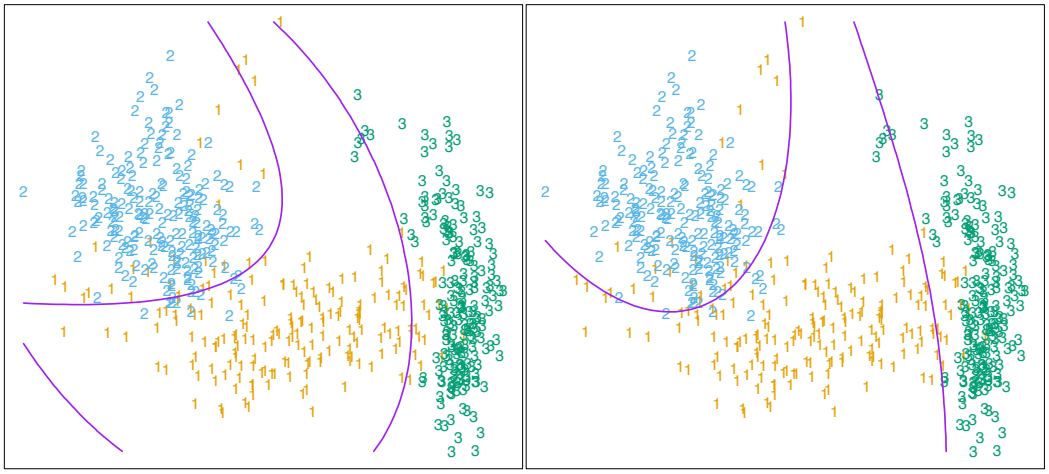
$$\arg\max_k \left[-log(2\pi)^{\frac{P}{2}}|\Sigma_k|^{\frac{1}{2}} - \frac{1}{2}(\mathbf{X}-\mu_k)'\Sigma_k^{-1}(\mathbf{X}-\mu_k) + log\pi_k \right]$$

- Decision boundary between classes k and l
  - $\{\mathbf{X}:\ \Pr\{Y=k|\mathbf{X}\} = \Pr\{Y=l|\mathbf{X}\}\}$

$$\begin{aligned} &= \{\mathbf{X}:\ \delta_k(\mathbf{X}) = \delta_l(\mathbf{X})\} \\ &= \{\mathbf{X}:\ log\delta_k(\mathbf{X}) - log\delta_l(\mathbf{X}) = 0\} \\ &= \left\{\mathbf{X}:\ log\frac{\pi_k}{\pi_l} - log\frac{|\Sigma_k|}{|\Sigma_l|} - \frac{1}{2}(\mathbf{X}-\mu_k)'\Sigma_k^{-1}(\mathbf{X}-\mu_k) \right. \\ &\qquad \left. + \frac{1}{2}(\mathbf{X}-\mu_l)'\Sigma_l^{-1}(\mathbf{X}-\mu_l) = 0 \right\} \end{aligned}$$

  - The quadratic terms do not cancel

Quadratic Discriminant Analysis



- Left: LDA in augmented five-dimensional space
- Right: QDA.
  - Both LDA and QDA typically perform well
  - Not because most data are Gaussian, or because most covariance matrices are equal
  - More likely, most data support simple decision boundaries, and the estimates via the Gaussian models are stable.
  - This is also a bias vs variance trade-off: estimate the boundary with bias, but less variance

Hastie, Tibshirani, Friedman, *The Elements of Statistical Learning* 2008

Estimation of the decision boundary

- Estimation from the training set:
  - $\widehat{\pi}_k = N_k/N$
  - $\widehat{\mu}_k = \sum_{i:Y_i=k} \mathbf{X}_i/N_k$
  - $\widehat{\mathbf{\Sigma}}_{P\times P} = \sum_{k=1}^{K} \sum_{i:Y_i=k} (\mathbf{X}_i - \widehat{\mu}_k)(\mathbf{X}_i - \widehat{\mu}_k)'/(N-K)$

- Haw many parameters in LDA?
  - Approx. $(K-1) \times (P+1)$ parameters
  - Only need the differences $\delta_k(\mathbf{X}) - \delta_K(\mathbf{X})$, where K is a pre-chosen class (e.g the last)
  - Each difference requires P + 1 parameters

- Haw many parameters in QDA?
  - Estimate separate covariance matrices per class
  - Large increase in parameters when P increases
  - Approx. $(K-1) \times \{P\,(P+3)/2+1\}$

Nearest Centroids: (diagonal-covariance LDA)

- Data x_{ij}
  - Observations $i = 1, \ldots, N$ and features $j = 1, \ldots, P$

- Training set centroid for feature j:

  in class k : $\bar{x}_{jk} = \sum_{i:i \in C_k} x_{ij}/N_k$ and overall $\bar{x}_j = \sum_{j=1}^{P} x_{ij}/N$

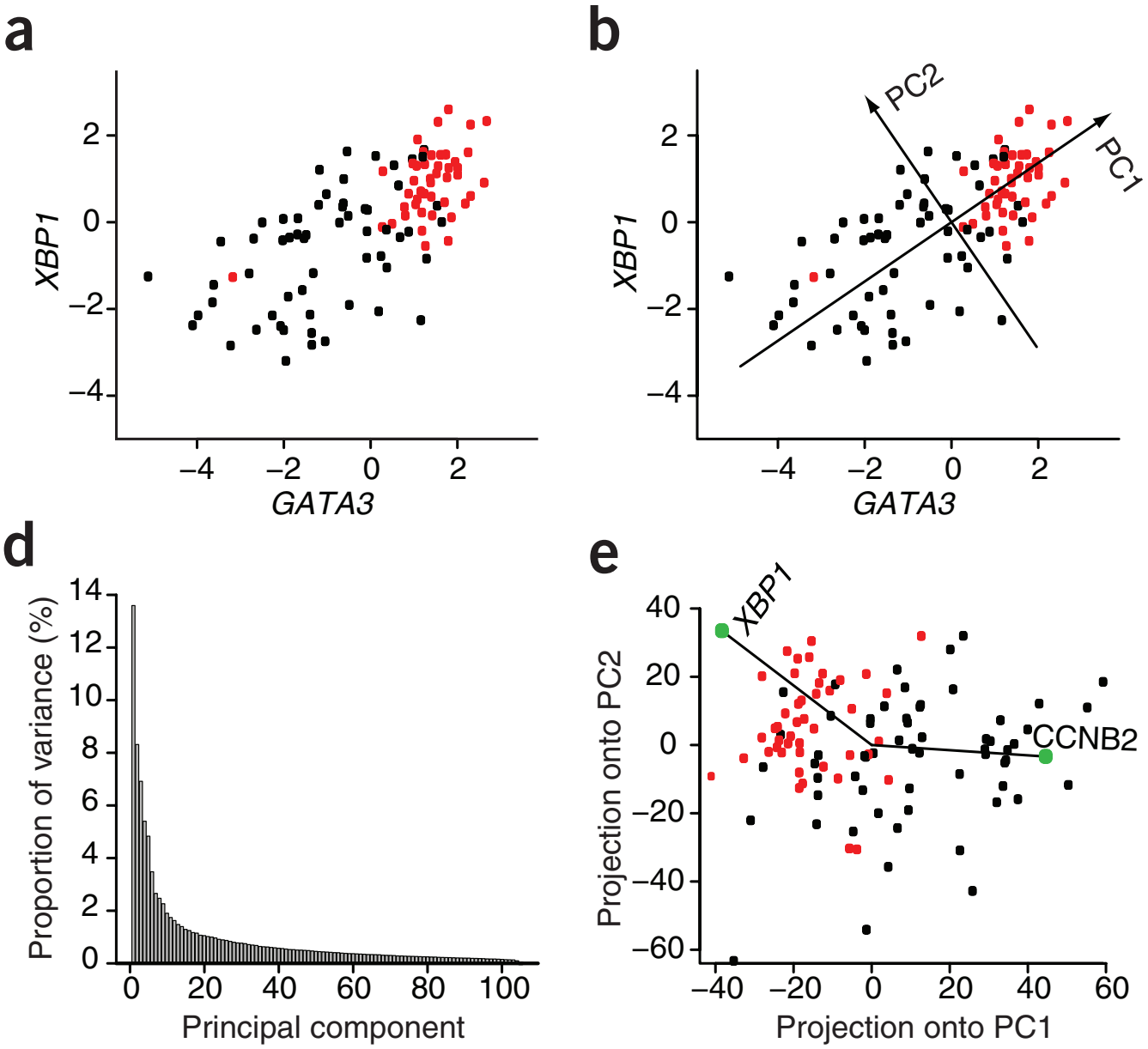
- Test set prediction
  - For $P >> N$, covariance estimation is unstable.
  - A simple solution: assume independent features (i.e. a diagonal Σ)
  - The discriminant score is

$$\delta_k(\mathbf{x}_i) = -\sum_{j=1}^{P} \frac{(x_{ij} - \bar{x}_{jk})^2}{s_j^2} + 2 \log \pi_k$$

  - The diagonal-covariance LDA classifier is equivalent to a nearest centroid classifier after standardization, correcting for class prior probability
  - $\hat{Y}(\mathbf{x}_i) = \arg \max_k \delta_k(\mathbf{x}_i)$

Connection to dimension reduction

Overview



Ringnér, Nature Biotechnology, 2008

Singular value decomposition

- Each sample as an observation in a G-dimensional space
  - Use the 'traditional' representation of the data (rows=observations; columns=variables)
  - X is the $I \times G$ matrix of centered feature expressions

	g_1	$\ldots$	g_P
S_1	$x_{11} - \bar{x}_{\cdot 1}$	$\ldots$	$x_{1P} - \bar{x}_{\cdot P}$
$\ldots$		$\ldots$	
S_N	$x_{N1} - \bar{x}_{\cdot 1}$	$\ldots$	$x_{NP} - \bar{x}_{\cdot P}$
$S_\cdot$	$\bar{x}_{\cdot 1}$	$\ldots$	$\bar{x}_{\cdot P}$

- Goal: find at most I linear combinations of features that best characterize the total between-sample variation

Singular value decomposition of X

- Each matrix can be represented with a Singular Value Decomposition

$$X = U \Lambda V'$$

  - U is a $N \times P$ orthonormal matrix (i.e. $U'U = I_P$)
    Columns of U are called left singular vectors

  - V is a $P \times P$ orthonormal matrix (i.e. $V'V = I_P$)
    Columns of U are called right singular vectors

  - Λ is a $P \times P$ diagonal matrix)
    The diagonal elements $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_P \geq 0$
    At most N λ_j are non-zero

- Right-multiply both sides by V:

$$X\ V = U\ \Lambda$$

  - The left side is the projection of the observations to unit vectors defined by columns of V

Singular value decomposition of X

- *Scores*
  - coordinates of the observations on the dimensions defined by columns of V
  - E.g. the coordinate of observation i in direction j is
    $\sum_{k=1}^{P} (x_{ik} - \bar{x}_{\cdot k}) \cdot v_{kj}$
  - Used to represent the observations in PCA plot

- *Loadings*
  - weight v_{kj} of feature k in the direction j
  - sometimes used to characterize the 'relative importance' of the feature in defining the direction
  - loadings are unique up to a sign flip (different software may return loadings with different sign)

- **Scores and loadings cannot be used as "measures of statistical significance"**

Principle component analysis of X

- Calculate the sample variance-covariance of matrix of features $\Sigma_{P \times P}$
  - The element (j, k) of the matrix is

$$\hat{\sigma}_{jk}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_{ij} - \bar{x}_{\cdot j}) \, (x_{ik} - \bar{x}_{\cdot k})$$

- Singular Value Decomposition of Σ:

$$(N-1)\Sigma = X' \, X = V \wedge U' \, U \wedge V' \overset{U \ orthonorm}{=} V \wedge^2 V'$$

  - Can obtain same left singular vectors V
  - The eigenvalues of covariance matrix $= \frac{1}{I-1} \lambda_j^2$

- SVD on the covariance matrix is referred to as Principle Component Analysis

Comments on PCA

- Estimation of the covariance of features is unstable, and SVD of X is preferred

- Decomposition of Σ helps interpretation

$$\sum_{j=1}^{P} \text{sample variance of columns of } X =$$

$$\sum_{j=1}^{P} \frac{1}{N-1} \sum_{i=1}^{N} (x_{ij} - \bar{x}_{\cdot j})^2 = \frac{1}{N-1} tr(X'X) =$$

$$\frac{1}{N-1} tr(V \, \Lambda^2 \, V') = \text{[property of trace]}$$

$$\frac{1}{N-1} tr(\Lambda^2 \, V' \, V) = \frac{1}{N-1} tr\Lambda^2 = \frac{1}{N-1} \sum_{j=1}^{P} \lambda_j^2$$

- The proportion of total variance explained by the principle component j is $\frac{\lambda_j^2}{\sum_k \lambda_k^2}$

- Select a subset of principle components that explain most variance of features between samples

Comments on PCA

- Problems with this approach
  - Does not distinguish useful (i.e. between-group) and nuisance (i.e. within-group) variation
  - Principle components can be driven by features with large nuisance variation
  - Cannot use scores as evidence of the ability of the features to predict the labels of the observations
- A typical modification
  - In addition to centering features, standardize them
  - Σ becomes the sample correlation matrix
  - Problem: standardization can remove useful variation of a feature between observations
- Conclusion
  - PCA performs best when all features have similar nuisance variation

Example: 2 features

- Sample correlation matrix of 2 features:

$$\Sigma = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$$

- Find the eigenvalues: Solve

$$det(\Sigma - \lambda I) = 0 \longrightarrow (1-\lambda)^2 - r^2 = 0$$

$\lambda_1 = 1 + r,\ \lambda_2 = 1 - r$; note that $\lambda_1 + \lambda_2 = 2 = tr(\Sigma)$

- Find the first eigenvector: Solve

$$\begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \begin{pmatrix} v_{11} \\ v_{21} \end{pmatrix} = \lambda_1 \begin{pmatrix} v_{11} \\ v_{21} \end{pmatrix}$$

- The equations are

$v_{11} + rv_{21} = (1+r)v_{11}$ and $rv_{11} + v_{21} = (1+r)v_{21}$

$\longrightarrow v_{11} = v_{21}$

- The orthonormality constraint

$$v_1' v_1 = v_{11}^2 + v_{21}^2 = 1$$

$$\longrightarrow v_{11} = v_{21} = \frac{1}{\sqrt{2}}$$

Example: 2 features

- Similarly find the second eigenvector
- This results in

$$V = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

- Coordinate of obs. i in 1rst direction

$$\sum_{k=1}^{2} \frac{x_{ik} - \bar{x}_{\cdot k}}{s_{\cdot k}} \cdot v_{kj} = \sum_{k=1}^{2} x'_{ik} \cdot v_{kj} = x'_{i1} \cdot \frac{1}{\sqrt{2}} + x'_{i2} \cdot \frac{1}{\sqrt{2}}$$

- Comments:
  - Coefficients (columns of V) independent of r
  - The variance in the new direction:

$$Var\{\frac{1}{\sqrt{2}}(x'_{i1} + x'_{i2})\} = \frac{1}{2} Var\{x'_{i1} + x'_{i2}\} =$$

$$\frac{1}{2}(Var\{x'_{i1}\} + Var\{x'_{i2}\} + 2Cov\{x'_{i1}, x'_{i2}\} = 1 + r$$

  - The larger r, the more proportion of the variance is accounted for by the **first** component
  - The larger r, the less proportion of the variance is accounted for by the **second** component

Summary:
SVD vs PCA (vs MDS)

- **SVD**: Singular Value Decomposition

$$X_{N\times P} = U_{N\times P}\,\Lambda_{P\times P}\,V'_{P\times P}$$

  - Decomposition of the original data matrix
  - Numerically stable, used for computation

- **PCA**: Principle Component Analysis

$$X'X_{P\times P} = V\Lambda U'U\Lambda V' = V_{P\times P}\,\Lambda^2_{P\times P}\,V'_{P\times P}$$

  - Numerically unstable, not used for computation
  - Useful to relate λ_j to % of explained variation

- **MDS**: Multidimensional Scaling

$$XX'_{N\times N} = U\Lambda V'V\Lambda U' = U_{N\times P}\,\Lambda^2_{P\times P}\,U'_{P\times N}$$

  - Used when raw data are in form of pairwise (dis-)similarities between observations

Inversion of Σ^{-1} in $\delta_k(\mathbf{X})$

- For LDA, the discriminant function is

  $$\delta_k(\mathbf{X}) = -\frac{1}{2}\log|\Sigma| - \frac{1}{2}(\mathbf{X}-\mu_k)'\Sigma^{-1}(\mathbf{X}-\mu_k) + log\pi_k$$

- Recall from PCA:

  $$\Sigma_{P\times P} = V_{P\times P}\Lambda_{P\times P}U'_{P\times N}U_{N\times P}\Lambda V'_{P\times P} = V_{P\times P}\,\Lambda^2_{P\times P}\,V'_{P\times P}$$

- Therefore

  $$(\mathbf{X}-\mu_k)'\Sigma^{-1}(\mathbf{X}-\mu_k) = (\mathbf{X}-\mu_k)'(V\Lambda^2V')^{-1}(\mathbf{X}-\mu_k)$$
  $$= \left[\Lambda^{-1}V'(\mathbf{X}-\mu_k)'\right]'\left[\Lambda^{-1}V'(\mathbf{X}-\mu_k)'\right]$$

- Equivalent to
  - Transform data: $\Lambda^{-1}V'\mathbf{X} \to X^*$, $\Lambda^{-1}V'\mu_k \to \mu^*$
  - The data are spheres on the new scale
  - On the new scale, classify $\mathbf{X}^*$ to the closest class centroid

- If the priors π_k are not equal
  - The decision boundary will shift
  - Will remain perpendicular to the line connecting the means in the transformed space

Fisher's approach

- Goal:
  - Project high-dimensional data onto a lower-dimensional space to best separate class labels, i.e.

$$\max \frac{\text{between-group variance}}{\text{within-group variance}} = \max_a \frac{a'\mathbf{B}a}{a'\Sigma a}$$

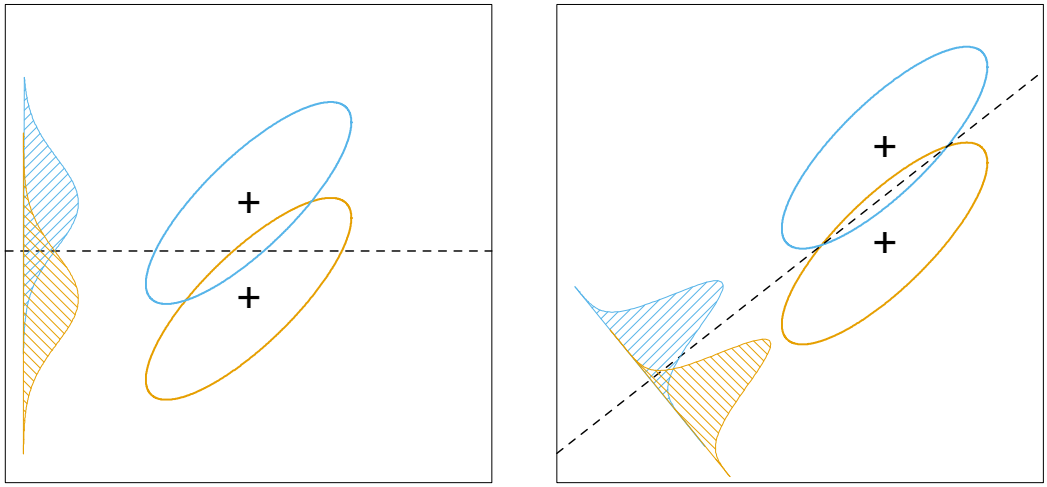
- Steps:
  - Transform the system of coordinates such that $a'\Sigma a = 1$ (with SVD/PCA)
  - Maximize $a'\mathbf{B}a$ in the transformed system of coordinates (with SVD/PCA)
  - Option: choose a subset of the transformed coordinates

- Decision for a new observation:
  - Assign the class with the closest center in the original variables

$$\begin{aligned} & \arg \min_k (\mathbf{X} - \mu_k)'\Sigma^{-1}(\mathbf{X} - \mu_k) \\ = \; & \arg \max_k -(\mathbf{X} - \mu_k)'\Sigma^{-1}(\mathbf{X} - \mu_k) \end{aligned}$$

  - Can show that maximizing this is equivalent to maximizing $\delta_k(\mathbf{x}_i)$

Optimal classification



- Left: Original coordinates
  - The line joining the centroids defines the direction of greatest centroid spread
  - The projected data overlap because of the covariance
- Right: Transformed coordinates
  - The discriminant direction minimizes this overlap for Gaussian data (right panel)

Hastie, Tibshirani, Friedman
The Elements of Statistical Learning 2008

Statistical regularization

Regularized LDA

- Shrink the separate covariances of QDA toward a common covariance as in LDA

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

- Shrink the common covariance in LDA toward a scalar covariance

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

- Again choose α by cross-validation

Nearest regular centroids

- Notation
  - x_{ij}: features $i = 1, \ldots, P$ in obs. $j = 1, \ldots, n$
  - each observation belongs to a class $k = 1, 2, \ldots, K$
  - C_k: indices of observations in class k
- Training set centroid for feature i

  in class k $\bar{x}_{ik} = \sum_{j \in C_k} x_{ij}/N_k$ and overall $\bar{x}_i = \sum_{j=1}^{n} x_{ij}/N_k$
- Standardization for feature i:

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{m_k \cdot s_i}$$

  $s_i^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{j \in C_k} (x_{ij} - \bar{x}_{ik})^2$ and $m_k = \sqrt{1/n_k - 1/n}$

  (the denumerator is the SE of the numerator)
  - higher weight to features with low variance within obs. of the same class
- Test set prediction
  - the class with the closest centroid

Nearest shrunken centroids

- Start with Nearest Centroids
  - d_{ik} is a t-statistic for feature i
  - compares class k to the average class

- Can re-write the standardized version as

$$\bar{x}_{ik} = \bar{x}_i + m_k \cdot s_i \cdot d_{ik}$$

- Propose to shrink d_{ik} towards 0

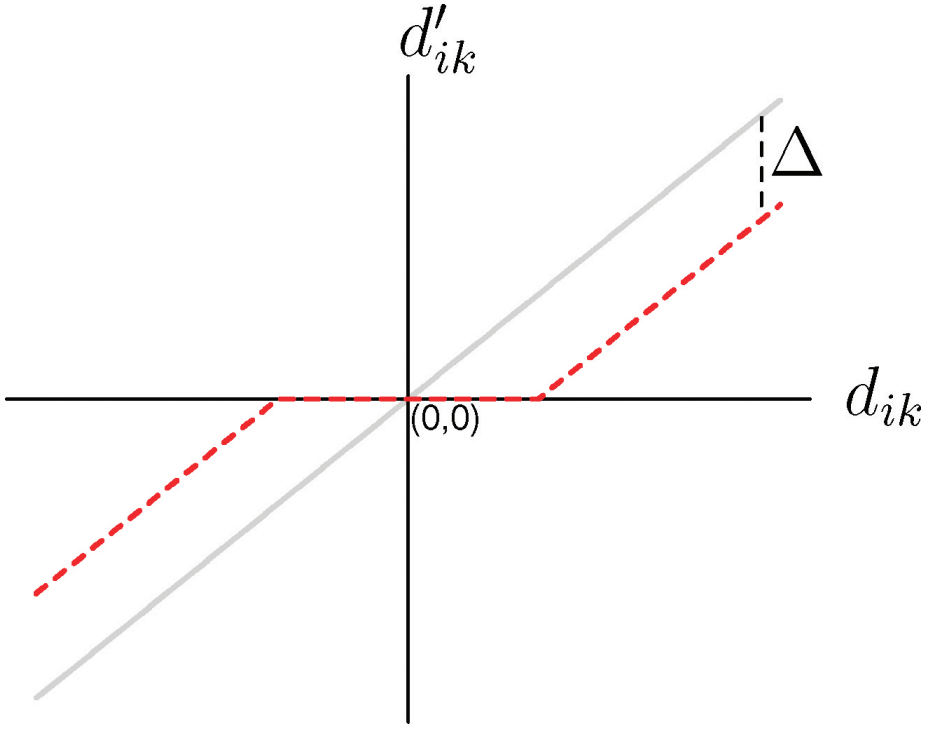
$$\bar{x}'_{ik} = \bar{x}_i + m_k \cdot s_i \cdot d'_{ik}$$

- Solution: soft thresholding

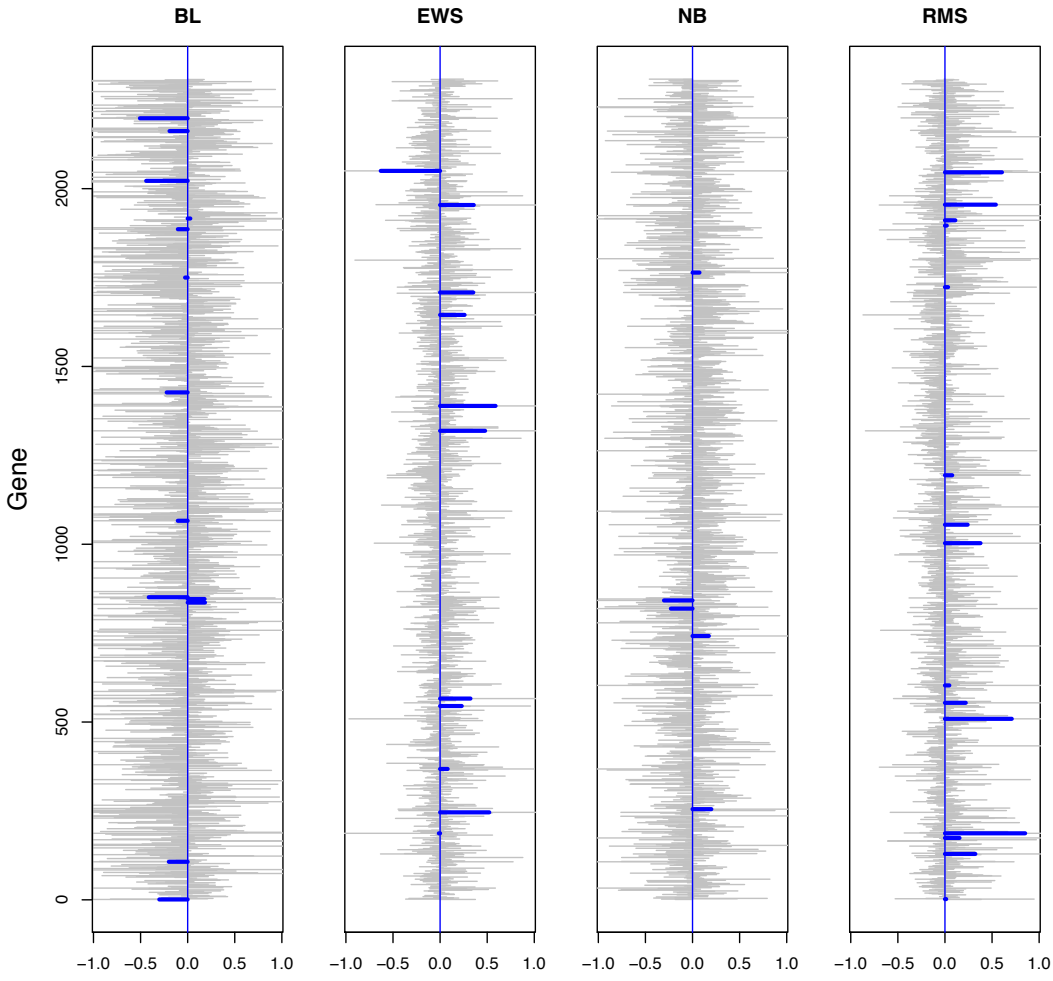
$$d\prime_{ik} = sign(d_{ik})(|d_{ik}| - \Delta)_+$$

  - where $t_+ = t$ if $t > 0$ and 0 otherwise
  - Δ chosen by cross-validation
  - Since many $\bar{x}_{ik}$ are noisy, soft thresholding produces more reliable estimates of the true means (Donoho and Johnstone, 1994)

Soft threshold function



Example

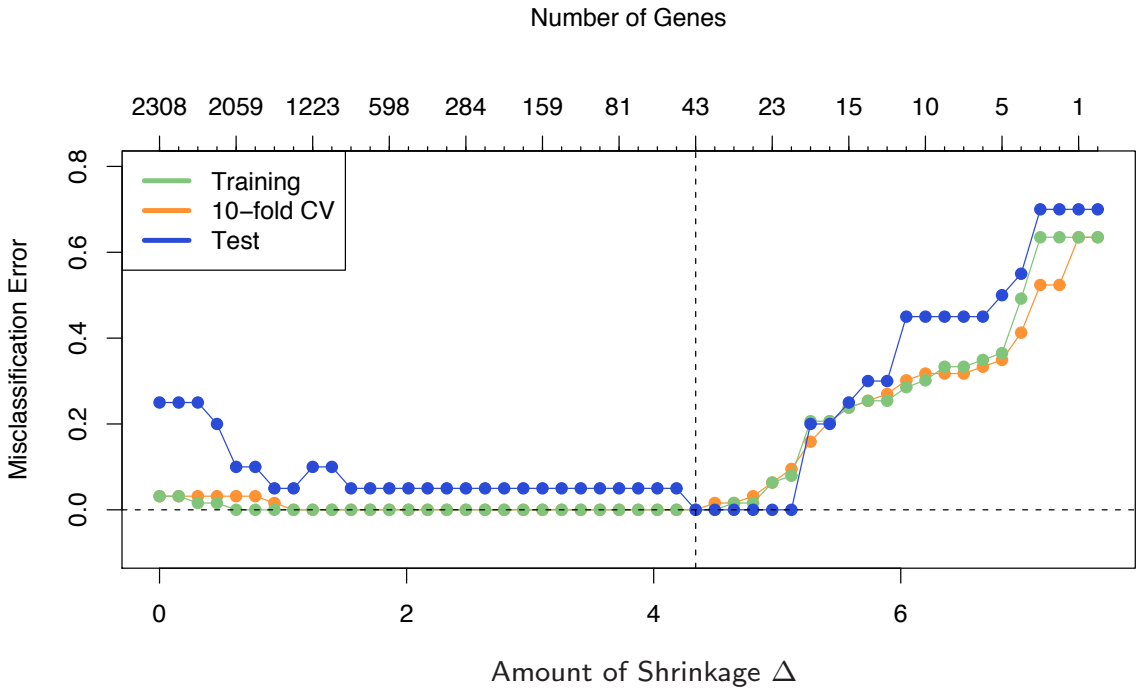


Centroids: Average Expression Centered at Overall Centroid

Training set of 144 patients with 14 different types of cancer, and a test set of 54 patients

Hastie, Tibshirani, Friedman, *The Elements of Statistical Learning* 2008

Example



Training set of 144 patients with 14 different types of cancer, and a test set of 54 patients

Hastie, Tibshirani, Friedman, *The Elements of Statistical Learning* 2008

Comments

- Choice of Δ
  - Choose max Δ (i.e. min features) for a same accuracy. Here $\Delta = 0.463$ yields 3,666 genes.
  - Different features are selected for a Δ at different folds of cross-validation.

- Properties of the classifier
  - d_{ik} accounts for the size of the class n_k in the denominator
    - Effectively applies a larger threshold to a smaller (higher variance) class
  - Some classes may be farther away than others from the overall centroid & easier to distinguish.
    - Many of the nonzero genes for that class may not be needed for accurate classification.
    - Adaptive class-specific thresholds improve the overall error rate (bottom panel).

Soft versus hard thresholding

- An alternative: hard thresholding
  - Keep all differences exceeding Δ in absolute value; discard the others
  - $d'_{ik} = d_{ik} \cdot I(|d_{ik}| > \Delta)$
  - I.e., differences $> \Delta$ are unchanged

- Not as advantageous
  - "jumpy" by nature: as Δ is increased, a gene with a full contribution d_{ik} is set to zero.
  - In simulations, soft thresholding yielded lower test error at its minimum

Class probabilities and discriminant functions

- Goals
  - Class probabilities in addition to classification
  - Adjust for class representation

- Notation
  - A test sample: $x^* = (x_1^*, x_2^*, \ldots, x_p^*)$

- Define the discriminant score for class k:

$$\delta_k(x^*) = \sum_{i=1}^{P} \frac{(x_i^* - \bar{x}'_{ik})^2}{s_i^2} - 2log\pi_k$$

  - $\delta_k(x^*)$ is the standardized squared distance of x^* to the kth shrunken centroid
  - $2log\pi_k$ is the correction based on the class prior probability π_k, $\sum_{k=1}^{K} \pi_k = 1$
  - π_k is the proportion of class k in the population
  - π_k gives preference to larger classes
  - $\widehat{\pi}_k = N_k/N$ (alternative: uniform or other priors)

Class probabilities and discriminant functions

- The classification rule:

$$C(x^*) = l \text{ if } \delta_l(x^*) = min_k \delta_k(x^*)$$

  - The prior plays an important role when multiple centroids are similarly close

- Class probabilities:

$$\widehat{p}_k(x^*) = \frac{exp\{-\frac{1}{2}\delta_k(x^*)\}}{\sum_{l=1}^{k} exp\{-\frac{1}{2}\delta_l(x^*)\}}$$

  - Analogy to Gaussian linear discriminant analysis

- Can use log-likelihood instead of cross-validation error rate to select Δ
  - With small # of samples & many classes, cross-validation curve can have discrete jumps and high variability
  - Mean cross-validated log-likelihood is typically smoother