Special Session on Pattern Classification

Maite Termenon[1]

[1]Computational Intelligence Group

2012 February 17

Book

Outline

Outline

Description

- An analytic, computationally feasible maximum likelihood solution lies in being able to find a parametric form for $p(x|\theta)$ that on the one hand matches the characteristics of the problem and on the other hand allows a reasonably tractable solution.

- There are distributions for which computationally feasible solutions can be obtained, and the key to their simplicity lies in the notion of a sufficient statistic.

Sufficient Statistics

- Sufficient statistic is a (possibly vector-valued) function $\mathbf{s}$ of the samples D that contains all of the information relevant to estimating some parameter θ.
- **Definition**: A statistic $\mathbf{s}$ is said to be sufficient for θ if $p(D \mid \mathbf{s}, \theta)$ is independent of θ.
- If we think of θ as a random variable, we can write:

$$p(\boldsymbol{\theta}|\mathbf{s}, \mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{s}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{s})}{p(\mathcal{D}|\mathbf{s})}, \qquad (56)$$

Sufficient Statistics

- It becomes evident that if $p(\boldsymbol{\theta}|\mathbf{s}, \mathcal{D}) = p(\boldsymbol{\theta}|\mathbf{s})$, $\mathbf{s}$ is sufficient for θ.
- Conversely, if $\mathbf{s}$ is a statistic for which $p(\boldsymbol{\theta}|\mathbf{s}, \mathcal{D}) = p(\boldsymbol{\theta}|\mathbf{s})$ and if $p(\boldsymbol{\theta}|\mathbf{s}) \neq 0$, then $p(\mathcal{D}|\mathbf{s}, \boldsymbol{\theta})$ is independent of θ.
- For a Gaussian distribution, the sample **mean** and **covariance**, together, represent a sufficient statistic for the true mean and covariance;

Factorization Theorem

- It states that **s** is sufficient for θ if and only if $p(D \mid \theta)$ can be factored into the product of two functions, one depending only on **s** and θ, and the other depending only on the training samples.

- It allows us to shift from the complicated density $p(D \mid s, \theta)$, used to define a sufficient statistic, to the simpler function:

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta}). \qquad (57)$$

- Characteristics of a sufficient statistic are completely determined by the density $p(x|\theta)$, and have nothing to do with a choice of an a priori density $p(\theta)$.

Factorization Theorem

Theorem 3.1 (Factorization) *A statistic* $\mathbf{s}$ *is sufficient for* $\boldsymbol{\theta}$ *if and only if the probability* $P(\mathcal{D}|\boldsymbol{\theta})$ *can be written as the product*

$$P(\mathcal{D}|\boldsymbol{\theta}) = g(\mathbf{s}, \boldsymbol{\theta})h(\mathcal{D}), \tag{58}$$

for some function $h(\cdot)$.

- The ability to factor $p(D \mid \theta)$ into a product $g(\mathbf{s}, \theta) h(D)$ is interesting only when the function g and the sufficient statistic s are simple.

- If s is a sufficient statistic for θ, this does not necessarily imply that their corresponding components are sufficient, i.e., that s_1 is sufficient for θ_1, or s_2 for θ_2, and so on.

Kernel Density

- The factoring of $p(D \mid \theta)$ into $g(\mathbf{s}, \theta) h(D)$ is not unique.
- If $f(\mathbf{s})$ is any function of s, then $g'(\mathbf{s}, \boldsymbol{\theta}) = f(\mathbf{s})g(\mathbf{s}, \boldsymbol{\theta})$ and $h'(D) = h(D)/f(\mathbf{s})$ are equivalent factors.
- This kind of ambiguity can be eliminated by defining the **kernel density**:

$$\bar{g}(\mathbf{s}, \boldsymbol{\theta}) = \frac{g(\mathbf{s}, \boldsymbol{\theta})}{\int g(\mathbf{s}, \boldsymbol{\theta}) \, d\boldsymbol{\theta}} \tag{63}$$

- which is invariant to this kind of scaling.

What for?

- What is the importance of sufficient statistics and kernel densities for parameter estimation?
  - For any clasification rule, we can find another based solely on sufficient statistics that has equal or better performance.
  - Data reduction: we can reduce an extremely large dataset down to a few numbers (sufficient statistics) confident that all relevant information has been preserved.

Estimators

- Bayes:
  - We can always create the Bayes classifier from sufficient statistics.
  - Our Bayes classifiers for Gaussian distributions were functions solely of the sufficient statistics, estimates of μ and $\sum$.

- Maximum likelihood:
  - When searching for a value of θ that maximizes $p(D \mid \theta) = g(\mathbf{s}, \theta) h(D)$, we can restrict our attention to $g(\mathbf{s}, \theta)$.
  - In this case, the normalization provided by kernel density is of no particular value unless $\bar{g}(\mathbf{s}, \theta)$ is simpler than $g(\mathbf{s}, \theta)$.

Kernel Density

- Significance of the kernel density is revealed in the Bayesian case.
- If we substitute $p(D \mid \theta) = g(\mathbf{s}, \theta) h(D)$ in Eq. 51, we obtain:

$$p(\theta | \mathcal{D}) = \frac{g(\mathbf{s}, \theta) p(\theta)}{\int g(\mathbf{s}, \theta) p(\theta) \, d\theta}. \qquad (64)$$

- If prior knowledge of θ is very vague, $p(\theta)$ will tend to be uniform, or changing very slowly as a function of θ.
- For a uniform $p(\theta)$, $p(\theta \mid D)$ is approximately the same as the kernel density.

Kernel Density

- The kernel density is the posterior distribution of the parameter vector when the prior distribution is uniform.
- When the a priori distribution is far from uniform, the kernel density typically gives the asymptotic distribution of the parameter vector.

Outline

Factorization Theorem

- To see how the Factorization Theorem can be used to obtain sufficient statistics, consider a familiar d-dimensional normal case with fixed covariance but unknown mean:

$$
\begin{aligned}
p(\mathcal{D}|\boldsymbol{\theta}) &= \prod_{k=1}^{n} \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}_k - \boldsymbol{\theta})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\theta})\right] \\
&= \frac{1}{(2\pi)^{nd/2}|\boldsymbol{\Sigma}|^{n/2}} \exp\left[-\frac{1}{2}\sum_{k=1}^{n}(\boldsymbol{\theta}^t \boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} - 2\boldsymbol{\theta}^t \boldsymbol{\Sigma}^{-1}\mathbf{x}_k + \mathbf{x}_k^t \boldsymbol{\Sigma}^{-1}\mathbf{x}_k)\right] \\
&= \exp\left[-\frac{n}{2}\boldsymbol{\theta}^t \boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} + \boldsymbol{\theta}^t \boldsymbol{\Sigma}^{-1}\left(\sum_{k=1}^{n}\mathbf{x}_k\right)\right] \\
&\quad \times \frac{1}{(2\pi)^{nd/2}|\boldsymbol{\Sigma}|^{n/2}} \exp\left[-\frac{1}{2}\sum_{k=1}^{n}\mathbf{x}_k^t \boldsymbol{\Sigma}^{-1}\mathbf{x}_k\right]. \qquad (65)
\end{aligned}
$$

Kernel Density

This factoring isolates the $\boldsymbol{\theta}$ dependence of $p(\mathcal{D}|\boldsymbol{\theta})$ in the first term, and hence from the Factorization Theorem we conclude that $\sum_{k=1}^{n} \mathbf{x}_k$ is sufficient for $\boldsymbol{\theta}$. Of course, any one-to-one function of this statistic is also sufficient for $\boldsymbol{\theta}$; in particular, the sample mean

$$\hat{\boldsymbol{\mu}}_n = \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k \tag{66}$$

is also sufficient for $\boldsymbol{\theta}$. Using this statistic, we can write

$$g(\hat{\boldsymbol{\mu}}_n, \boldsymbol{\theta}) = \exp\left[-\frac{n}{2}\left(\boldsymbol{\theta}^t \boldsymbol{\Sigma}^{-1} \boldsymbol{\theta} - 2\boldsymbol{\theta}^t \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\mu}}_n\right)\right]. \tag{67}$$

From using Eq. 63, or by completing the square, we can obtain the kernel density:

$$\bar{g}(\hat{\boldsymbol{\mu}}_n, \boldsymbol{\theta}) = \frac{1}{(2\pi)^{d/2}|\frac{1}{n}\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\mu}}_n)^t \left(\frac{1}{n}\boldsymbol{\Sigma}\right)^{-1} (\boldsymbol{\theta} - \hat{\boldsymbol{\mu}}_n)\right]. \tag{68}$$

These results make it immediately clear that $\hat{\boldsymbol{\mu}}_n$ is the maximum likelihood estimate for $\boldsymbol{\theta}$.

Sufficient Statistics for Exponential Family

This same general approach can be used to find sufficient statistics for other density functions. In particular, it applies to any member of the *exponential family*, a group of probability and probability density functions that possess simple sufficient statistics. Members of the exponential family include the Gaussian, exponential, Rayleigh, Poisson, and many other familiar distributions. They can all be written in the form

$$p(\mathbf{x}|\boldsymbol{\theta}) = \alpha(\mathbf{x}) \exp \left[\mathbf{a}(\boldsymbol{\theta}) + \mathbf{b}(\boldsymbol{\theta})^t \mathbf{c}(\mathbf{x}) \right]. \tag{69}$$

If we multiply n terms of the form in Eq. 69 we find

$$p(\mathcal{D}|\boldsymbol{\theta}) = \exp \left[n\mathbf{a}(\boldsymbol{\theta}) + \mathbf{b}(\boldsymbol{\theta})^t \sum_{k=1}^{n} \mathbf{c}(\mathbf{x}_k) \right] \prod_{k=1}^{n} \alpha(\mathbf{x}_k) = g(\mathbf{s}, \boldsymbol{\theta}) h(\mathcal{D}), \tag{70}$$

where we can take

$$
\begin{aligned}
\mathbf{s} &= \frac{1}{n} \sum_{k=1}^{n} \mathbf{c}(\mathbf{x}), \\
g(\mathbf{s}, \boldsymbol{\theta}) &= \exp \left[n\{ \mathbf{a}(\boldsymbol{\theta}) + \mathbf{b}(\boldsymbol{\theta})^t \mathbf{s} \} \right],
\end{aligned}
$$

and

$$h(\mathcal{D}) = \prod_{k=1}^{n} \alpha(\mathbf{x}_k).$$

Distributions, Sufficient Statistics and Unnormalized Kernels

Table 3.1: Common Exponential Distributions and their Sufficient Statistics.

Name	Distribution	Domain		s	$[g(\mathbf{s},\boldsymbol{\theta})]^{1/n}$
Normal	$p(x\vert\boldsymbol{\theta}) = \sqrt{\frac{\theta_2}{2\pi}} e^{-(1/2)\theta_2(x-\theta_1)^2}$	$\theta_2 > 0$		$\begin{bmatrix} \frac{1}{n}\sum_{k=1}^{n} x_k \\ \frac{1}{n}\sum_{k=1}^{n} x_k^2 \end{bmatrix}$	$\sqrt{\theta_2}\, e^{-\frac{1}{2}\theta_2(s_2 - 2\theta_1 s_1 + \theta_1^2)}$
Multi-variate Normal	$p(\mathbf{x}\vert\boldsymbol{\theta}) = \frac{\lvert\Theta_2\rvert^{1/2}}{(2\pi)^{d/2}} e^{-(1/2)(\mathbf{x}-\boldsymbol{\theta}_1)^t \Theta_2 (\mathbf{x}-\boldsymbol{\theta}_1)}$	Θ_2 positive definite		$\begin{bmatrix} \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k \\ \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k \mathbf{x}_k^t \end{bmatrix}$	$\lvert\Theta_2\rvert^{1/2}\, e^{-\frac{1}{2}[\mathrm{tr}\Theta_2 s_2 - 2\theta_1^t \Theta_2 s_1 + \theta_1^t \Theta_2 \theta_1]}$
Exponential	$p(x\vert\theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$\theta > 0$		$\frac{1}{n}\sum_{k=1}^{n} x_k$	$\theta e^{-\theta s}$
Rayleigh	$p(x\vert\theta) = \begin{cases} 2\theta x e^{-\theta x^2} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$\theta > 0$		$\frac{1}{n}\sum_{k=1}^{n} x_k^2$	$\theta e^{-\theta s}$
Maxwell	$p(x\vert\theta) = \begin{cases} \frac{4}{\sqrt{\pi}}\theta^{3/2} x^2 e^{-\theta x^2} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$\theta > 0$		$\frac{1}{n}\sum_{k=1}^{n} x_k^2$	$\theta^{3/2} e^{-\theta s}$
Gamma	$p(x\vert\boldsymbol{\theta}) = \begin{cases} \frac{\theta_2^{\theta_1+1}}{\Gamma(\theta_1+1)} x^{\theta_1} e^{-\theta_2 x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$	$\theta_1 > -1$ $\theta_2 > 0$		$\begin{bmatrix} \left(\prod_{k=1}^{n} x_k\right)^{1/n} \\ \frac{1}{n}\sum_{k=1}^{n} x_k \end{bmatrix}$	$\frac{\theta_2^{\theta_1+1}}{\Gamma(\theta_1+1)} s_1^{\theta_1} e^{-\theta_2 s_2}$
Beta	$p(x\vert\boldsymbol{\theta}) = \begin{cases} \frac{\Gamma(\theta_1+\theta_2+2)}{\Gamma(\theta_1+1)\Gamma(\theta_2+1)} x^{\theta_1}(1-x)^{\theta_2} & 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$	$\theta_1 > -1$ $\theta_2 > -1$		$\begin{bmatrix} \left(\prod_{k=1}^{n} x_k\right)^{1/n} \\ \left(\prod_{k=1}^{n} (1-x_k)\right)^{1/n} \end{bmatrix}$	$\frac{\Gamma(\theta_1+\theta_2+2)}{\Gamma(\theta_1+1)\Gamma(\theta_2+1)} s_1^{\theta_1} s_2^{\theta_2}$

Outline

Two Issues

- There are two issues that must be confronted:
  - How classification accuracy depends upon the dimensionality (and amount of training data).
  - Computational complexity of designing the classifier.

Outline

Accuracy, Dimension and Training Sample Size

- Consider the two-class multivariate normal case with the same covariance where $p(\mathbf{x}|\omega_j) \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}), j = 1, 2$. If the a priori probabilities are equal, then Bayes error rate is:

$$P(e) = \frac{1}{\sqrt{2\pi}} \int\limits_{r/2}^{\infty} e^{-u^2/2} \, du, \tag{71}$$

where r^2 is the squared Mahalanobis distance (Chap. ??, Sect. ??):

$$r^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \tag{72}$$

Thus, the probability of error decreases as r increases, approaching zero as r approaches infinity. In the conditionally independent case, $\boldsymbol{\Sigma} = diag(\sigma_1^2, ..., \sigma_d^2)$, and

$$r^2 = \sum_{i=1}^{d} \left(\frac{\mu_{i1} - \mu_{i2}}{\sigma_i} \right)^2. \tag{73}$$

- This shows how each feature contributes to reducing the probability of error.

Useful Features

- Most useful features are the ones for which the difference between the means is large relative to the standard deviations.
- No feature is useless if its means for the two classes differ.
- A way to reduce the error rate further is to introduce new, independent features.
- Although increasing the number of features increases the cost and complexity of both the feature extractor and the classifier, it is often reasonable to believe that the performance will improve.
- In practice, beyond a certain point, the inclusion of additional features leads to worse rather than better performance.

Example
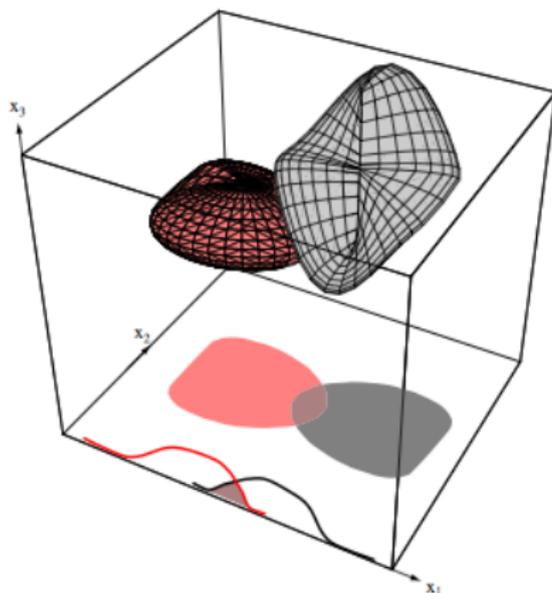
Figure 3.3: Two three-dimensional distributions have nonoverlapping densities, and thus in three dimensions the Bayes error vanishes. When projected to a subspace — here, the two-dimensional $x_1 - x_2$ subspace or a one-dimensional x_1 subspace — there can be greater overlap of the projected distributions, and hence greater Bayes errors.

Outline

1. 3.6 *Suficient Statistics
   - Sufficient Statistics and the Exponential Family

2. 3.7 Problems of Dimensionality
   - Accuracy, Dimension and Training Sample Size
   - Computational Complexity
   - Overfitting

Order of a Function

- $f(x)$ is "of the order of $h(x)$" - written $f(x) = O(h(x))$" - if there exist constants c_0 and x_0 such that $| f(x) | \leq c_0 | h(x) |$ for all $x > x_0$.

- This means that for sufficiently large x, an upper bound on the function grows no worse than $h(x)$.

Computational complexity of an algorithm

- We are generally interested in the number of basic mathematical operations (additions, multiplications and divisions) it requires, or in the time and memory needed on a computer.

- To illustrare this concept, we consider the complexity of a maximum likelihood estimation of the parameters in a classifier for Gaussian priors in d dimensions, with n training samples for each of c categories.

- For each category it is necessary to calculate the discriminant function of:

$$g(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \overset{O(dn)}{\overset{\uparrow}{\hat{\boldsymbol{\mu}}}})^t \overbrace{\widehat{\boldsymbol{\Sigma}}^{-1}}^{O(nd^2)} (\mathbf{x} - \hat{\boldsymbol{\mu}}) - \overbrace{\frac{d}{2}}^{O(1)} \ln 2\pi - \overbrace{\frac{1}{2} \ln |\widehat{\boldsymbol{\Sigma}}|}^{O(d^2n)} + \overbrace{\ln P(\omega)}^{O(n)}. \quad (74)$$

Computational complexity of an algorithm

- We assume that $n > d$ (otherwise our covariance matrix will not have a well defined inverse).

- For large problems, the overall complexity of calculating an individual discriminant function is dominated by the $O(d^2 n)$ term.

Estimating the covariance matrix

- This requires the estimation of $d(d+1)/2$ parameters:
    - d diagonal elements.
    - $d(d-1)/2$ independent off-diagonal elements.
- Maximum likelihood estimate:

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{k=1}^{n} (\mathbf{x}_k - \mathbf{m}_n)(\mathbf{x}_k - \mathbf{m}_n)^t, \qquad (75)$$

- is the sum of $n-1$ independent d - by - d matrices of rank one, and thus is guaranteed to be singular if $n \leq d$.
- Since we must invert $\widehat{\boldsymbol{\Sigma}}$ to obtain the discriminant functions, we have an algebraic requirement for at least $d+1$ samples.

Computational complexity for classification

- Computational complexity for classification is less.
- Given a test point x we must compute $(x - \mu)$, an $O(d)$ calculation.
- For each of the categories, we must multiply the inverse covariance matrix by the separation vector, an $O(d^2)$ calculation.
- The $max_i g_i(x)$ decision is a separate $O(c)$ operation. For small c then, recall is an $O(d^2)$ operation.
- Recall is much simpler (and faster) than learning.

Space and Time Complexities

- For instance, the sample mean of a category could be calculated with d separate processors, each adding n sample values.
- We can describe it as:
    - $O(d)$ in space
    - $O(n)$ in time
- For any particular algorithm, there may be a number of time-space tradeoffs, for instance using a single processor many times, or using many processors in parallel for a shorter time.
- Such tradeoffs are important considerations.

Qualitative Distinctions

- Distinction is made between *polynomially* complex and *exponentially* complex algorithms - $O(a^k)$ for some constant a and aspect or variable k of the problem.

- Exponential algorithms are generally so complex that for reasonable size cases we avoid them altogether, and approximate solutions that can be found by polynomially complex algorithms.

Outline

Overfitting

- Number of available samples is inadequate, and the question of how to proceed arises.
- Possibilities:
  - To reduce the dimensionality by:
    - redesigning the feature extractor,
    - selecting an appropriate subset of the existing features,
    - combining the existing features in some way.
  - To assume that all c classes share the same covariance matrix, and to pool the available data.
  - To look for a better estimate for Σ.
  - To assume statistical independence

Paradox

- The classifier that results from assuming independence is almost certainly suboptimal.

- It will perform better if it happens that the features actually are independent.

- But... how can it provide better performance when this assumption is untrue?

- The answer again involves the problem of **insufficient data**, and some insight into its nature can be gained from considering an analogous problem in curve fitting.

Curve Fitting
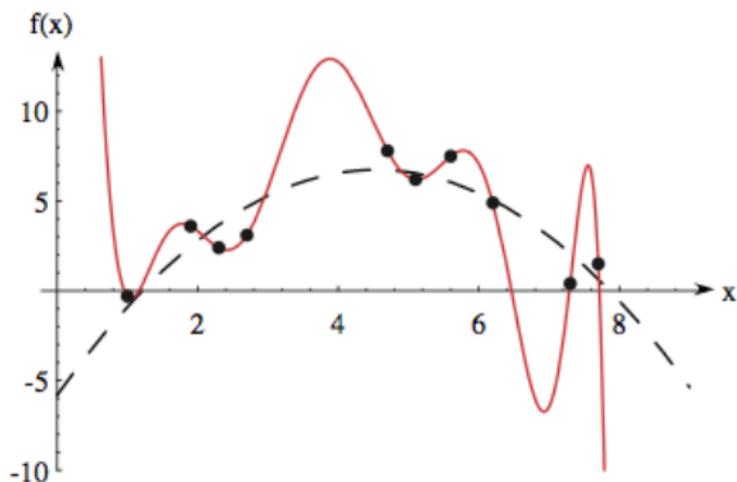
Figure 3.4: The "training data" (black dots) were selected from a quadradic function plus Gaussian noise, i.e., $f(x) = ax^2 + bx + c + \epsilon$ where $p(\epsilon) \sim N(0, \sigma^2)$. The 10th degree polynomial shown fits the data perfectly, but we desire instead the second-order function $f(x)$, since it would lead to better predictions for new samples.

Improving generalization

- We might consider beginning with a high-order polynomial (e.g., 10th order), and successively smoothing or simplifying our model by eliminating the highest-order terms.

- Heuristic methods that can be applied in the Gaussian classifier case.

  - We wish to design a classifier for distributions $N(\mu_1, \Sigma_1)$ and $N(\mu_2, \Sigma_2)$ and we have insufficient data for accurately estimating the parameters.

  - We might make the simplification that they have the same covariance, i.e., $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$, and estimate Σ accordingly (such estimation requires proper normalization of the data).

Improving generalization II

- Intermediate approach: to assume a weighted combination of the equal and individual covariances, (known as shrinkage or regularized discriminant analysis) since the individual covariances "shrink" toward a common one.

If i is an index on the c categories in question, we have

$$\Sigma_i(\alpha) = \frac{(1-\alpha)n_i\Sigma_i + \alpha n \Sigma}{(1-\alpha)n_i + \alpha n}, \qquad (76)$$

for $0 < \alpha < 1$. Additionally, we could "shrink" the estimate of the (assumed) common covariance matrix toward the identity matrix, as

$$\Sigma(\beta) = (1-\beta)\Sigma + \beta\mathbf{I}, \qquad (77)$$

for $0 < \beta < 1$ (Computer exercise 8). (Such methods for simplifying classifiers have counterparts in regression, generally known as *ridge regression*.)