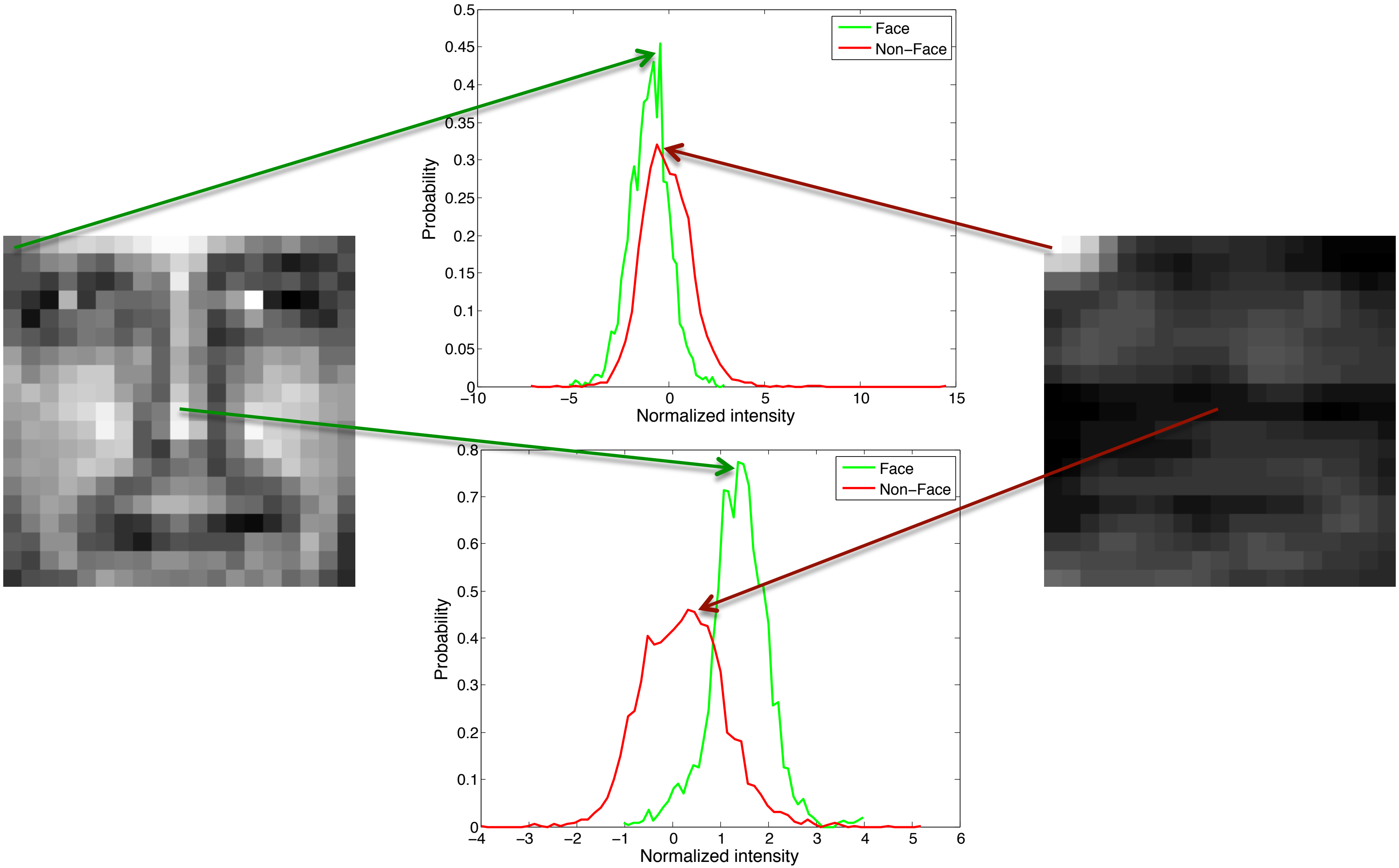
FEATURE SELECTION

J. Elder

Outline

- Analyzing individual features
  - Signal to noise ratio
  - Null hypothesis testing
- Analyzing feature vectors
  - Divergence
  - Chernoff Bound and Bhattacharyya Distance
  - Scatter Matrices
  - Feature Subset Selection

Which feature would you choose?



Outline

- Analyzing individual features
  - **Signal to noise ratio**
  - Null hypothesis testing
- Analyzing feature vectors
  - Divergence
  - Chernoff Bound and Bhattacharyya Distance
  - Scatter Matrices
  - Feature Subset Selection

Signal to Noise Ratio

- We wish to select features that have a high signal-to-noise ratio.
- Typically we measure the signal strength as the difference in the class-conditional means:

  $S = \mu_2 - \mu_1$

- The noise can be characterized by the conditional standard deviation. If the class-conditional distributions can be assumed to have the same standard deviation, the signal-to-noise ratio can be expressed as:

$$SNR = \frac{\mu_2 - \mu_1}{\sigma}$$

Signal to Noise

$$SNR = \frac{\mu_2 - \mu_1}{\sigma}$$

- Why is this a meaningful number?
- Suppose that the class-conditional distributions are Gaussian, i.e.,

$$x_1 \sim \mathcal{N}\left(\mu_1, \sigma^2\right), \qquad x_2 \sim \mathcal{N}\left(\mu_2, \sigma^2\right)$$

- Recall that for the equal variance case, the maximum likelihood classifier will select the class based upon the Mahalanobis distance. In this case, this means that

$$\begin{aligned} \left|x - \mu_1\right| < \left|x - \mu_2\right| \rightarrow \omega_1 \\ \left|x - \mu_1\right| > \left|x - \mu_2\right| \rightarrow \omega_2 \end{aligned}$$

- wlog, suppose $\mu_1 < \mu_2$. Then the decision rule becomes

$$\begin{aligned} x < x_0 \rightarrow \omega_1 \\ x > x_0 \rightarrow \omega_2 \end{aligned} \qquad \text{where } x_0 = \frac{1}{2}\left(\mu_1 + \mu_2\right)$$

Signal to Noise

- The probability of error is therefore:

$$p(\text{error}) = p(x > x_0 \mid \omega_1)p(\omega_1) + p(x < x_0 \mid \omega_2)p(\omega_2)$$

Let $F(x;\mu,\sigma^2)$ represent the cumulative normal distribution, i.e., $F(x;\mu,\sigma^2) \triangleq \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x} \exp\left(-\frac{(x'-\mu)^2}{2\sigma^2}\right)dx'$

Then $p(\text{error}) = \frac{1}{2}\left(1 - F(x_0;\mu_1,\sigma^2) + F(x_0;\mu_2,\sigma^2)\right)$

$$= \frac{1}{2}\left(1 - F\left(\frac{x_0-\mu_1}{\sigma};0,1\right) + F\left(\frac{x_0-\mu_2}{\sigma};0,1\right)\right)$$

$$= \frac{1}{2}\left(F\left(-\frac{x_0-\mu_1}{\sigma};0,1\right) + F\left(\frac{x_0-\mu_2}{\sigma};0,1\right)\right)$$

$$= \frac{1}{2}\left(F\left(-\frac{\mu_2-\mu_1}{2\sigma};0,1\right) + F\left(-\frac{\mu_2-\mu_1}{2\sigma};0,1\right)\right)$$

$$= F\left(-\frac{\mu_2-\mu_1}{2\sigma};0,1\right)$$

Thus the probability of error depends only on the SNR, $\frac{\mu_2-\mu_1}{\sigma}$.

SNR: Unknown Parameters

- Thus when selecting features a good first-order rule is to select the features with maximal SNR (for minimal error).
- Normally, we do not know the class-conditional parameters, and must estimate them from data.
- We could compute an ML estimate. However, convention is to use unbiased estimates of the parameters:

$$SNR = \frac{\bar{x}_2 - \bar{x}_1}{s}$$

where

$$\bar{x}_i = \frac{1}{N_i}\sum_{\omega_j = i} x_j \quad \text{and} \quad s^2 = \frac{\sum_{\omega_j=1}\left(x_j - \bar{x}_1\right)^2 + \sum_{\omega_j=2}\left(x_j - \bar{x}_2\right)^2}{N_1 + N_2 - 2}.$$

SNR: Unknown Parameters

$$SNR = \frac{\bar{x}_2 - \bar{x}_1}{s}$$

where

$$\bar{x}_i = \frac{1}{N_i} \sum_{\omega_j = i} x_j \quad \text{and} \quad s^2 = \frac{\sum_{\omega_j=1} \left(x_j - \bar{x}_1\right)^2 + \sum_{\omega_j=2} \left(x_j - \bar{x}_2\right)^2}{N_1 + N_2 - 2}.$$

- Note that due to sampling error, the estimated parameters will not be exactly correct.
- As a consequence, even if a feature is completely uninformative for the classification, the SNR estimate will still be non-zero.
- How can we filter out these uninformative features?

Outline

- Analyzing individual features
  - Signal to noise ratio
  - **Null hypothesis testing**
- Analyzing feature vectors
  - Divergence
  - Chernoff Bound and Bhattacharyya Distance
  - Scatter Matrices
  - Feature Subset Selection

Null Hypothesis Testing

$$SNR = \frac{\bar{x}_2 - \bar{x}_1}{s}.$$

- One approach is to use a null hypothesis testing (NHT) procedure.

  Suppose that the observations x are normally distributed.

  Since the $\bar{x}_i$ are linear functions of x, these are also normally distributed:

  $\bar{x}_i \sim \mathcal{N}\left(\mu_i, \sigma^2 / N_i\right)$ (Lecture 1 – Topic 7.2)

  Thus $\bar{x}_2 - \bar{x}_1$ is also normally distributed:

  $$\bar{x}_2 - \bar{x}_1 \sim \mathcal{N}\left(\mu_2 - \mu_1, \frac{\sigma^2}{N_1} + \frac{\sigma^2}{N_2}\right)$$

- Notice that the precision of this distribution increases linearly with the number of training inputs. In particular, if we knew the variance, we could form the test statistic

  $$z = \frac{\bar{x}_2 - \bar{x}_1}{\sigma\left(\frac{1}{N_1} + \frac{1}{N_2}\right)} \sim \mathcal{N}(0,1) \text{ if } \mu_1 = \mu_2.$$

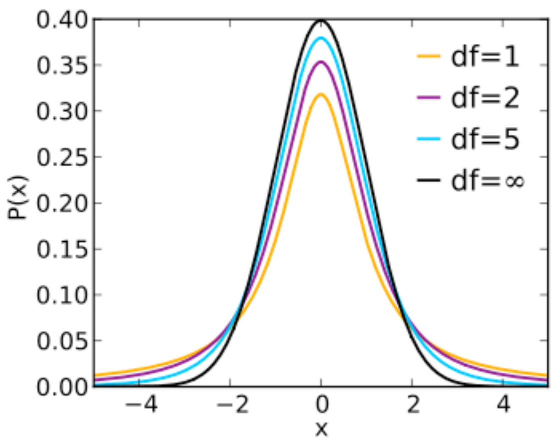
The t-statistic

$$SNR = \frac{\bar{x}_2 - \bar{x}_1}{s}.$$

- However, *s* is also a random variable, and so the SNR is not distributed as a Gaussian.
- Instead, under the null hypothesis (no difference in the means), the statistic

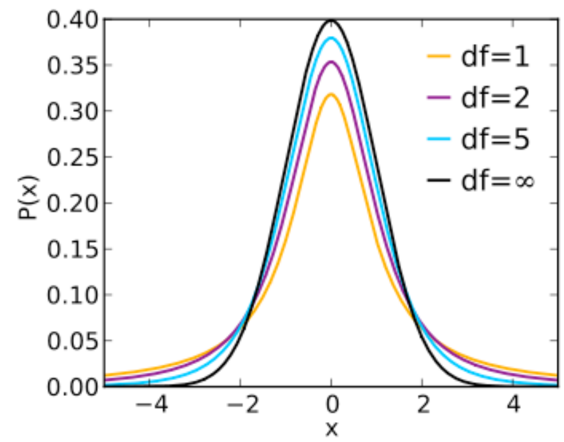
$$t = \frac{\bar{x}_2 - \bar{x}_1}{s\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$

follows a student's *t* distribution with N_1+N_2-2 degrees of freedom.



The t-test

$$t = \frac{\bar{x}_2 - \bar{x}_1}{s\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}$$



- Given this statistic, one can compute the probability that a value this large (in absolute value) would be produced were there no difference in the means (the null hypothesis).
- In the NHT procedure, we 'fail to reject' the null hypothesis if this probability exceeds a pre-specified value, typically .05.
- In selecting features, we can choose to reject any feature that does not meet this NHT criterion.

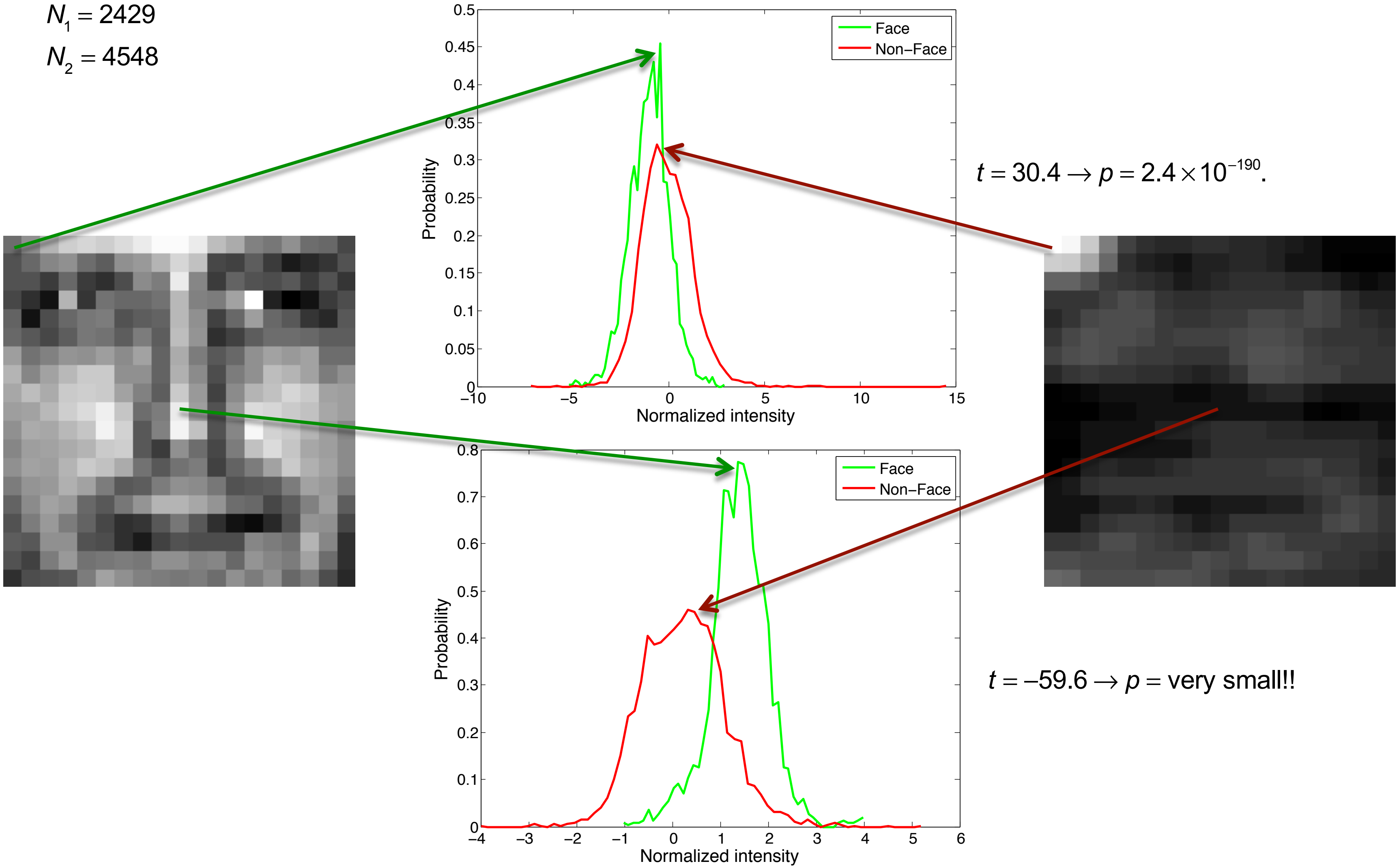
Generalizing to Multiple Classes

- The NHT approach can be generalized to K>2 classes using analysis of variance (ANOVA) methods.
- We will not cover these here.

Example

$N_1 = 2429$

$N_2 = 4548$



$t = 30.4 \rightarrow p = 2.4 \times 10^{-190}.$

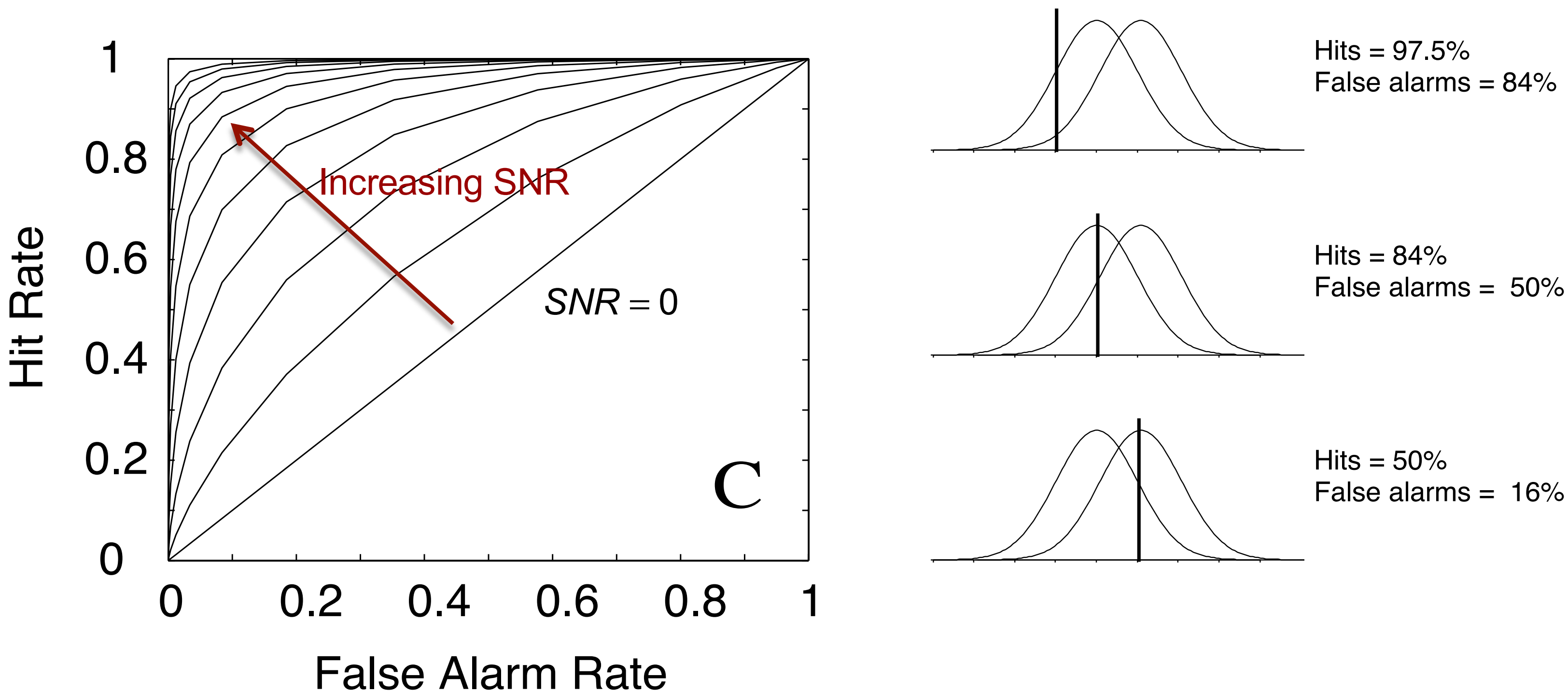
$t = -59.6 \rightarrow p = \text{very small!!}$

Limitations of NHT for Feature Selection

- A significant t-statistic indicates that there is sufficient training data to reveal a discriminative signal in a particular feature.
- However, it does not guarantee that including the feature will improve your classification rate on new data.
- The main problem is that the discriminative information in that feature may be redundant with information in other features.

ROC Curves

- As the criterion threshold is swept from right to left, p(HIT) increases, but p(FA) also increases.
- The resulting plot of p(HIT) vs p(FA) is called a receiver-operating characteristic (ROC).



End of Lecture

Outline

- Analyzing individual features
  - Signal to noise ratio
  - Null hypothesis testing
- **Analyzing feature vectors**
  - Divergence
  - Chernoff Bound and Bhattacharyya Distance
  - Scatter Matrices
  - Feature Subset Selection

Divergence

- The divergence between two class-conditional distributions is a symmetrization of the Kullback-Leibler distance between the two distributions:

Divergence $d_{ij} = D_{ij} + D_{ji}$

where $D_{ij} + D_{ji}$ are the Kullback-Leibler divergences:

$$D_{ij} = \int_{-\infty}^{\infty} p(\mathbf{x} \mid \omega_i) \log \frac{p(\mathbf{x} \mid \omega_i)}{p(\mathbf{x} \mid \omega_j)} d\mathbf{x} \; P(\omega_i)$$

$$D_{ji} = \int_{-\infty}^{\infty} p(\mathbf{x} \mid \omega_j) \log \frac{p(\mathbf{x} \mid \omega_j)}{p(\mathbf{x} \mid \omega_i)} d\mathbf{x}$$

- The separability of M classes can then be defined as

$$d = \sum_{i=1}^{M} \sum_{j=1}^{M} d_{ij} P(\omega_i) P(\omega_j)$$

Properties of the Divergence

- If the components of the feature vectors are conditionally independent, then

$$d_{ij}(\mathbf{x}) = \sum_{m=1}^{M} d_{ij}(x_m)$$

- The divergence is general in the sense that it can be non-zero even if the class-conditional means are the same.
- However, for multivariate normal distributions with equal covariance matrices, the divergence reduces to the Mahalanobis distance between the means:

$$d_{ij} = (\mu_i - \mu_j)^t \Sigma^{-1} (\mu_i - \mu_j)$$

$$d = \sum_{i=1}^{M} \sum_{j=1}^{M} d_{ij} P(\omega_i) P(\omega_j)$$

where $d_{ij} = D_{ij} + D_{ji}$

and

$$D_{ij} = \int_{-\infty}^{\infty} p(\mathbf{x} \mid \omega_i) \log \frac{p(\mathbf{x} \mid \omega_i)}{p(\mathbf{x} \mid \omega_j)} d\mathbf{x}$$

$$D_{ji} = \int_{-\infty}^{\infty} p(\mathbf{x} \mid \omega_j) \log \frac{p(\mathbf{x} \mid \omega_j)}{p(\mathbf{x} \mid \omega_i)} d\mathbf{x}$$

Chernoff Bound

- The minimum classification error P_e attainable by a Bayes classifier for two classes is given by

$$P_e = \int_{-\infty}^{\infty} \min\left(P(\omega_i)p(\mathbf{x} \mid \omega_i), P(\omega_j)p(\mathbf{x} \mid \omega_j)\right) d\mathbf{x}.$$

- Using the inequality

$$\min(a,b) \le a^s b^{1-s} \qquad \text{for } a,b \ge 0 \quad \text{and } 0 \le s \le 1$$

yields the Chernoff bound:

$$P_e \le P(\omega_i)^s P(\omega_j)^{1-s} \int_{-\infty}^{\infty} p(\mathbf{x} \mid \omega_i)^s \, p(\mathbf{x} \mid \omega_j)^{1-s} \, d\mathbf{x}$$

- Since this must apply for all s between 0 and 1, one can (in theory) find the tightest bound by constrained minimization with respect to s.

Bhattacharyya Distance

- Using $s = 1/2$, and assuming multivariate normal distributions, a specific form of the Chernoff distance known as the Bhattacharyya distance can be derived.
- Again, if the covariances are equal, the Bhattacharyya distance is proportional to the Mahalanobis distance.

Scatter Matrices

- The divergence and Chernoff bound are only readily computed for normal distributions.
- A more easily computed measure of class separability is based upon scatter matrices.
- The within-class scatter matrix S_w is defined as:

$$S_w = \sum_{i=1}^{M} P_i \Sigma_i,$$

where $P_i = \frac{n_i}{N}$ is the empirical estimate of the prior for Class ω_i

and Σ_i is the covariance matrix for Class ω_i

- The between-class scatter matrix S_b is defined as:

$$S_b = \sum_{i=1}^{M} P_i \left(\mu_i - \mu_0\right)\left(\mu_i - \mu_0\right)^t,$$

where μ_i is the mean for Class ω_i and μ_0 is the pooled mean.

- The mixture scatter matrix S_m is defined as:

$$S_m = E\left[\left(\mathbf{x} - \mu_0\right)\left(\mathbf{x} - \mu_0\right)^t\right].$$

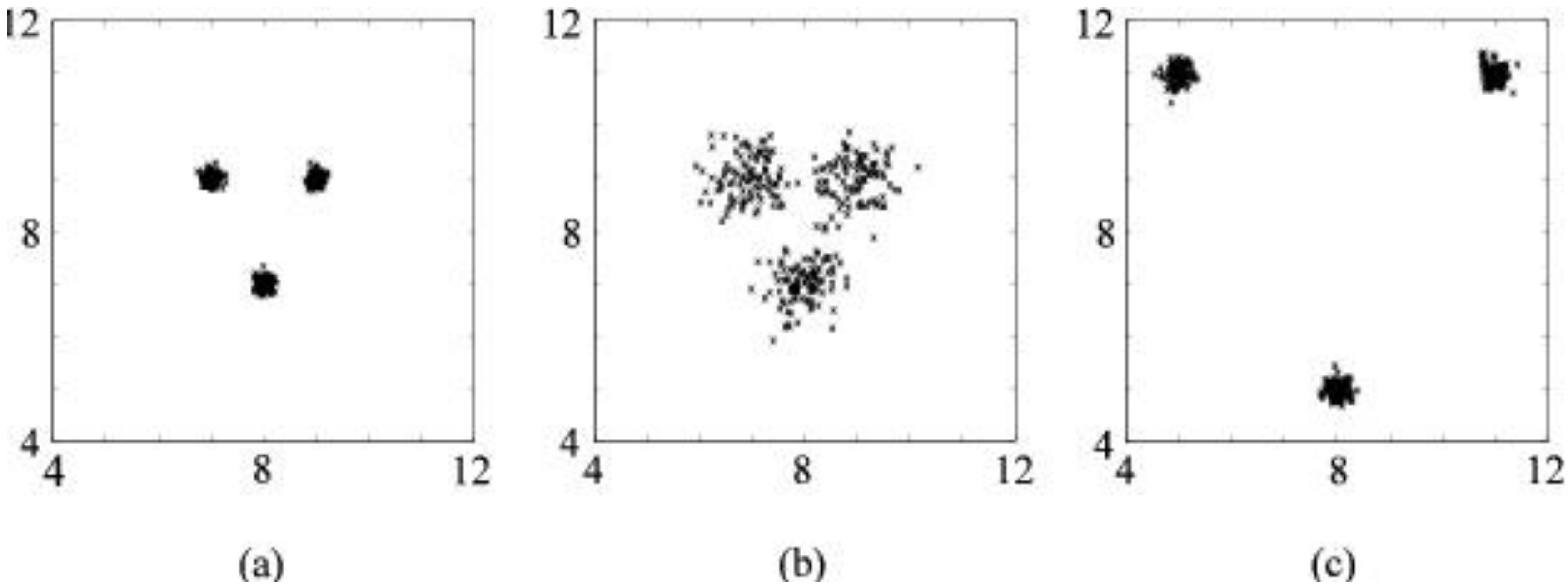
Scatter Matrices

- Measures of separability can be formed from these scatter matrices. Recalling that:
  - The trace of the covariance matrix is equal to the sum of the eigenalues and is a measure of the total variance in the data
  - The determinant of the covariance matrix can also be used as a measure of the total variability and is sometimes called the generalized variance
- We have the following measures of separability:

$$J_1 = \frac{\text{trace}(S_m)}{\text{trace}(S_w)} \qquad J_2 = \frac{|S_m|}{|S_w|} \qquad J_3 = \frac{|S_b|}{|S_w|}$$

Example

$J_3 = 164.7$ $J_3 = 12.5$ $J_3 = 620.9$



Selecting Feature Vectors

- The methods we have discussed provide a means for selecting a single feature.
- One can use any of these to select a subset of features based upon their individual merits.
- However, this does not take into account the statistical redundancy between these features.
- Ch 5.7.2 of the textbook discusses some heuristics for reducing this redundancy when selecting feature subsets. We will not discuss these here.
- We will discuss more powerful and principled methods for selecting feature subsets (dimensionality reduction, boosting) in coming lectures.