CSci 5521 **Fall 2020**

Handout: Bayes with Gaussian Distribution v2

Bayes Theory

Goal: Given a set of classes $c_1, \ldots, c_n$ and a sample with a given set of feature values which we denote by a vector $\mathbf{x}$, the goal is to determine which class the sample is most likely to belong to. The choice is made by selecting that class yielding the largest of the posterior probabilities that the class is c_i given that the features are $\mathbf{x}$. This posterior probability is denoted

$$p(\text{class} = c_i \,|\, \text{feature} = \mathbf{x}) \tag{1}$$

1. But we don't have these posterior probabilities. Instead we have the conditional distribution functions describing how feature values are distributed within each class, $p(\text{feature} = \mathbf{x} \,|\, \text{class} = c_i)$, and the relative probability for each class (the prior) $p(\text{class} = c_i)$. This information may come from either (a) a-priori theory or (b) derived from a large training set of samples.

2. So we use Bayes Theory to derive the posterior probabilities (1):

$$\begin{aligned} p(\text{class} = c_i \,|\, \text{feature} = \mathbf{x}) &= \frac{p(\text{class} = c_i \;\&\; \text{feature} = \mathbf{x})}{p(\text{feature} = \mathbf{x})} \\ &= p(\text{feature} = \mathbf{x} \,|\, \text{class} = c_i) \cdot \frac{p(\text{class} = c_i)}{p(\text{feature} = \mathbf{x})} \\ &= \text{conditional probability} \cdot \frac{\text{prior}}{\text{evidence}} \end{aligned} \tag{2}$$

   Then our *decision method* is to choose the most likely class given a particular value for the feature, i.e. the class c_{i^*} for which the posterior probability (2) is the maximum over all possible classes. Since the denominator remains the same, depending only on the sample, this the same class we obtain by maximizing the *joint probability*, $p(\text{class} = c_i \;\&\; \text{feature} = \mathbf{x})$ for a fixed $\mathbf{x}$. Hence it is sufficient to maximize the joint probability.

3. Though often the distributions within each class, $p(\text{feature} = \mathbf{x} \,|\, \text{class} = c_i)$, obeys a standard distribution function, no such assumption has been made for the derivation of (2) here. It may be that the feature vector $\mathbf{x}$ resides in a discrete space containing only finitely many possible values, in which case the distributions, and hence the joint probabilities, can be defined by a table lookup.

4. If we have only two classes, then we can define a *discriminant function* which is positive if the sample is classified in class c_2 and negative if the sample is classified in c_1. Two possibilities for such a discriminant function are

$$\begin{aligned} & d_1(\mathbf{x}) = p(\text{class} = c_2 \,|\, \text{feature} = \mathbf{x}) - p(\text{class} = c_1 \,|\, \text{feature} = \mathbf{x}), \\ \text{or} \quad & d_2(\mathbf{x}) = \log\left(\frac{p(\text{class} = c_2 \,|\, \text{feature} = \mathbf{x})}{p(\text{class} = c_1 \,|\, \text{feature} = \mathbf{x})} \right). \end{aligned} \tag{4}$$

Normal Distributions.

Now suppose that we have two classes c_1 and c_2, and the feature values are normally distributed within each class. That is:

$$p(\text{feature} = \mathbf{x} \,|\, \text{class} = c_1) = \frac{1}{\sqrt{(2\pi)^d \cdot |\Sigma_1|}} \exp\left(\frac{(\mathbf{x} - \boldsymbol{\mu}_1)^T \Sigma_1^{-1} (\mathbf{x} - \boldsymbol{\mu}_1)}{-2} \right) \tag{5}$$

where $\boldsymbol{\mu}_1$ is the mean of all the class c_1 and Σ_1 is the corresponding covariance matrix. Here $|\Sigma|$ denotes the determinant of Σ. The other class c_2 also obeys a similar distribution function

$$p(\text{feature} = \mathbf{x} \,|\, \text{class} = c_2) = \frac{1}{\sqrt{(2\pi)^d \cdot |\Sigma_2|}} \exp\left(\frac{(\mathbf{x}-\boldsymbol{\mu}_2)^T \Sigma_2^{-1} (\mathbf{x}-\boldsymbol{\mu}_2)}{-2} \right) \tag{6}$$

We can then form the discriminant function $d_2(\mathbf{x})$ (4) to obtain a way of choosing the class for any given set of feature values $\mathbf{x}$. We get that

$$d_2(\mathbf{x}) = \log \left[\frac{\dfrac{p(c_2)}{p(\mathbf{x})\sqrt{(2\pi)^d \cdot |\Sigma_2|}} \exp\left(\dfrac{(\mathbf{x}-\boldsymbol{\mu}_2)^T \Sigma_2^{-1} (\mathbf{x}-\boldsymbol{\mu}_2)}{-2} \right)}{\dfrac{p(c_1)}{p(\mathbf{x})\sqrt{(2\pi)^d \cdot |\Sigma_1|}} \exp\left(\dfrac{(\mathbf{x}-\boldsymbol{\mu}_1)^T \Sigma_1^{-1} (\mathbf{x}-\boldsymbol{\mu}_1)}{-2} \right)} \right], \tag{7}$$

where $p(c_i)$, $p(\mathbf{x})$ are shorthand for $p(\text{class} = c_i)$, $p(\text{feature} = \mathbf{x})$, respectively.

We apply the "log" function, cancelling several terms, to arrive at

$$d_2(\mathbf{x}) = \frac{1}{2} \left[\left((\mathbf{x}-\boldsymbol{\mu}_1)^T \Sigma_1^{-1} (\mathbf{x}-\boldsymbol{\mu}_1) \right) - \left((\mathbf{x}-\boldsymbol{\mu}_2)^T \Sigma_2^{-1} (\mathbf{x}-\boldsymbol{\mu}_2) \right) + \log \frac{|\Sigma_1|}{|\Sigma_2|} \right] + \log \frac{p(c_2)}{p(c_1)} \tag{8}$$

The terms in $\mathbf{x}$ can be further simplified to

$$\begin{aligned} d_2(\mathbf{x}) = \quad & \frac{1}{2} \mathbf{x}^T \left(\Sigma_1^{-1} - \Sigma_2^{-1} \right) \mathbf{x} + \left(\Sigma_2^{-1} \boldsymbol{\mu}_2 - \Sigma_1^{-1} \boldsymbol{\mu}_1 \right)^T \mathbf{x} \\ & + \frac{1}{2} \left(\boldsymbol{\mu}_1^T \Sigma_1^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^T \Sigma_2^{-1} \boldsymbol{\mu}_2 \right) + \frac{1}{2} \log \frac{|\Sigma_1|}{|\Sigma_2|} - \log \frac{p(c_1)}{p(c_2)} \end{aligned} \tag{9}$$

Example - Covariance Matrices the Same

A common example occurs if $\Sigma_1 = \Sigma_2$. In this case we just use Σ to denote the common matrix. Then the quadratic and some of the constant terms in (9) drop out, leaving us with

$$\begin{aligned} d_2(\mathbf{x}) \quad &= \left(\Sigma^{-1} \boldsymbol{\mu}_2 - \Sigma^{-1} \boldsymbol{\mu}_1 \right)^T \mathbf{x} + \left\{ \frac{1}{2} \left(\boldsymbol{\mu}_1^T \Sigma^{-1} \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^T \Sigma^{-1} \boldsymbol{\mu}_2 \right) + \log \frac{p(c_2)}{p(c_1)} \right\} \\ &= \left(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 \right)^T \Sigma^{-1} \mathbf{x} + \left\{ \frac{1}{2} \left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \right)^T \Sigma^{-1} \left(\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1 \right) - \log \frac{p(c_1)}{p(c_2)} \right\} \\ &= \left(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 \right)^T \Sigma^{-1} \left(\mathbf{x} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2} \right) + \log \frac{p(c_2)}{p(c_1)} \,. \end{aligned} \tag{10}$$

Thus $d_2(\mathbf{x})$ is a *linear discriminant function* of the form $\mathbf{a}^T(\mathbf{x}-\mathbf{v}) - C$, with direction $\mathbf{a} = \Sigma^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$, $\mathbf{v} = (\boldsymbol{\mu}_2 + \boldsymbol{\mu}_1)/2$ (the midpoint between the two centers), and $C = [\log p(c_1) / p(c_2)]$. The last line in (10) shows that cut between the 2 classes is the hyper-plane normal to $\mathbf{a}$, passing through the point $\mathbf{v}$ halfway between the two means, but shifted by the log of the *class likelihood function*: $C = \log[p(c_1) / p(c_2)]$.

Numerical Example - 1D

You are an admissions officer at a college examining applications for admission. Suppose the feature is the applicant's ACT score, and class c_1 is the class of students who graduate in 5 years. The ACT score are normally distributed, using the corresponding scalar version of the formulas (5) and (6), for $i = 1, 2$:

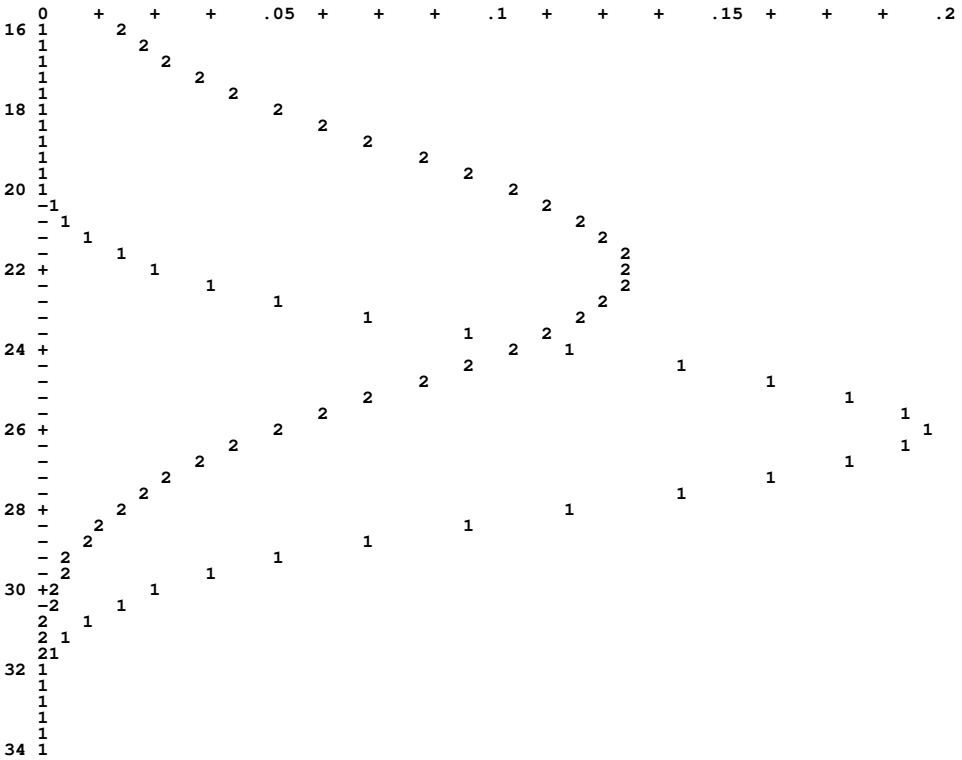
$$p(\text{ACT} = x \,|\, \text{class} = c_i) = \frac{1}{\sigma_i \cdot \sqrt{2\pi}} \exp\left(\frac{(x-\mu_i)^2}{-2\sigma_i^2}\right), \tag{11}$$

where

item	c_1	c_2
$\mu_i =$	26	22
$\sigma_i =$	2	3

Look at these cheap sketches sideways...

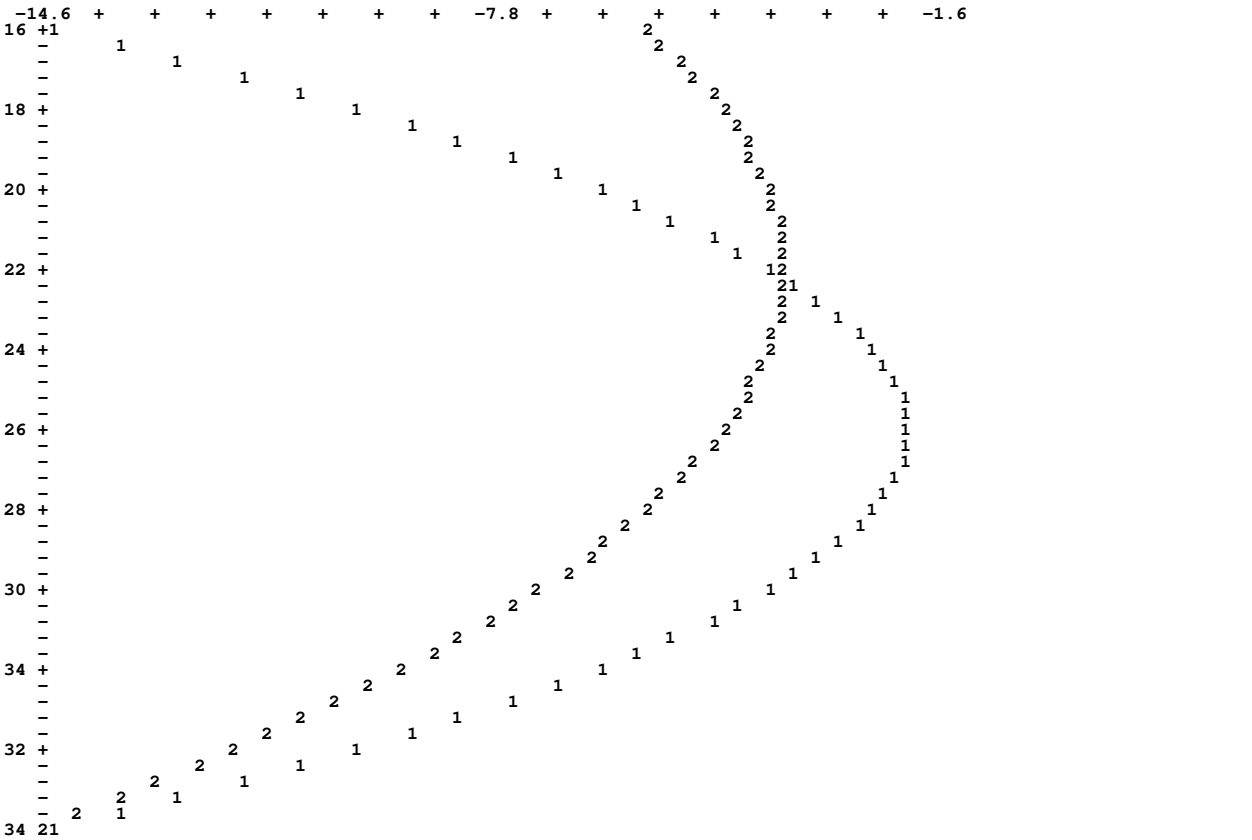
class conditional probabilities: vertical axis is abscissa (ACT value)



Here the two covariances are different, so we must use the general form of the discriminant function (9), but simplified for the scalar case. We also need the values of the prior probabilities. In this case we assume the priors are $p(c_1) = 0.8$ and $p(c_2) = 0.2$. Then the discriminant function function (9) reduces to

$$\begin{aligned} d_2(x) &= \frac{1}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right)x^2 + \left(\frac{\mu_2}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2}\right)x + \frac{1}{2}\left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2}\right) + \frac{1}{2}\log\frac{\sigma_1^2}{\sigma_2^2} - \log\frac{p(c_1)}{p(c_2)} \\ &= \frac{1}{2}x^2\left(\frac{1}{4} - \frac{1}{9}\right) + \left(\frac{22}{9} - \frac{26}{4}\right)x + \frac{1}{2}\left(\frac{26^2}{4} - \frac{22^2}{9}\right) + \frac{1}{2}\log\frac{4}{9} - \log\frac{0.8}{0.2} \qquad (12) \\ &= .069444444x^2 - 4.0555555x + 55.8193516, \end{aligned}$$

with roots 36.1886576, 22.2113423. In between the two roots, d_2 is negative, indicating the choice of class c_1. Beyond either root, d_2 is positive, indicating the choice c_2.



Numerical Example - 2D - Same Covariance Matrix

Suppose we also use class percentile (y) as well as the ACT scores (x), using the following parameters:

item	c_1	c_2
prior	0.8	0.2
μ_x	26	22
μ_y	85	70
σ_x	2	
σ_y	5	
σ_{xy}	6	

Then the parameters in the normal distribution function (6) are

$$\mu_1 = \begin{bmatrix} 26 \\ 85 \end{bmatrix}, \mu_2 = \begin{bmatrix} 22 \\ 70 \end{bmatrix}, \Sigma = \begin{bmatrix} 4 & 6 \\ 6 & 25 \end{bmatrix}. \tag{13}$$

Since we are using the same covariance for both classes, we can use the linear discriminant function (10) of the form $\mathbf{a}^T(\mathbf{x} - \mathbf{v}) - C$ with

$$\mathbf{a} = \begin{bmatrix} 4 & 6 \\ 6 & 25 \end{bmatrix}^{-1}\left(\begin{bmatrix} 22 \\ 70 \end{bmatrix} - \begin{bmatrix} 26 \\ 85 \end{bmatrix}\right) = \begin{bmatrix} 4 & 6 \\ 6 & 25 \end{bmatrix}^{-1}\begin{bmatrix} -4 \\ -15 \end{bmatrix} = \begin{bmatrix} -.15625 \\ -.56250 \end{bmatrix},$$

$$\mathbf{v} = \frac{1}{2}\left(\begin{bmatrix} 26 \\ 85 \end{bmatrix} + \begin{bmatrix} 22 \\ 70 \end{bmatrix}\right) = \begin{bmatrix} 24.0 \\ 77.5 \end{bmatrix},$$

$$C = \log\frac{0.8}{0.2} = 1.3862943.$$

Here's a sketch showing one level curve for the two joint probability density functions. The two functions are surfaces with the same shape, but the surface for class 2 has been shrunk relative to class 1 by the priors.

level curves and linear separator for two 2D normal distributions