Ollscoil Teicneolaíochta na Mumhan
Munster Technological University

Machine Learning

Lecture 8: Bayesian classification (theory)

Bayesian model

- The basic assumption of Bayesian models for classification is that every feature of an object is a realisation of a random variable
- The distribution of these random variables depends only on the class of the object

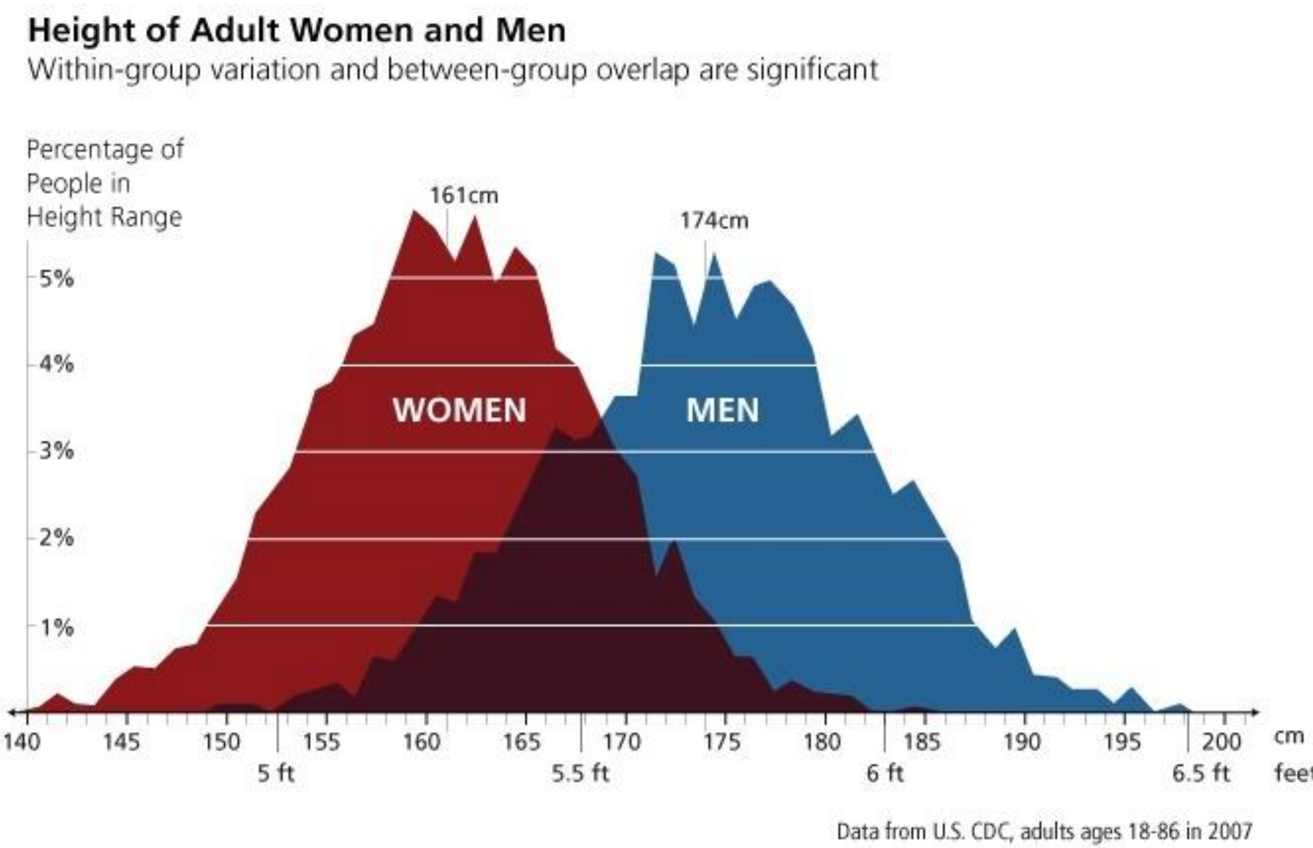
- Let's assume there are a total of k different possible classes of objects $\{\omega_1, \dots, \omega_k\}$
- We now observe an object of class ω_i and extract the feature vector $x \in \mathbb{R}^d$
- Then observations of this feature vector happen with a certain probability given by the class-specific **likelihood**

$$P[x|\omega_i]$$

Likelihood

- The likelihood $P[x|\omega_i]$ tells us for each class, how likely it is to observe a feature of a member of this class

- Example: $P[height|women]$ and $P[height|men]$



Likelihood

- The likelihood $P[x|\omega_i]$ tells us for each class, how likely it is to observe a feature of a member of this class

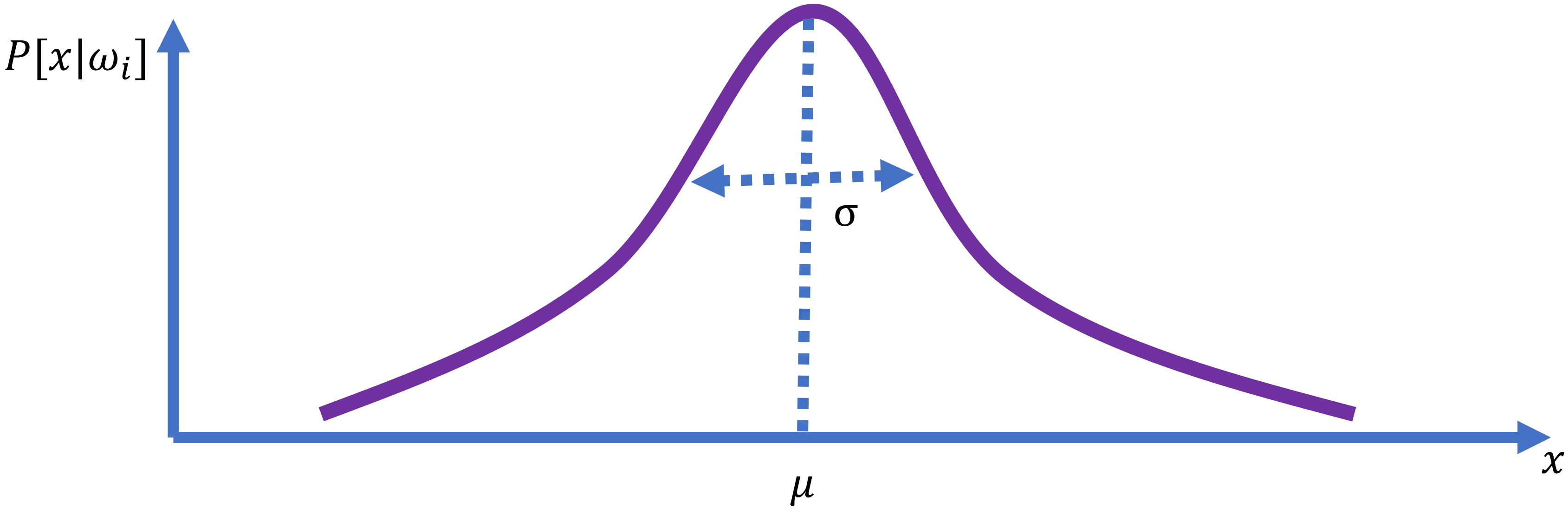
- Example:
  - For the classes “English words” and “French words” the probability of a particular letter occurring in these words is different

  - For instance the likelihood of observing the letter “H” is
$$P["H"|"English"] = 4.96\%$$
$$P["\mathrm{H}"|"\mathrm{French}"] = 0.93\%$$
  - While likelihood of observing the letter “U” is
$$P["U"|"English"] = 2.68\%$$
$$P["\mathrm{U}"|"\mathrm{French}"] = 5.55\%$$

Likelihood

- The likelihood $P[x|\omega_i]$ tells us for each class, how likely it is to observe a feature of a member of this class

- For discrete features the likelihood can be determined by counting occurrence for each class (e.g. letters in words)
- For continuous features the likelihood can be parameterised, often as Normal distribution with mean μ and variance σ^2, and then estimated from data (e.g. height distribution in population)



Bayes' Theorem

- The likelihood $P[x|\omega_i]$ tells us for each class, how likely it is to observe a feature of a member of this class

- Our problem is the opposite, though: We are measuring a feature and want to determine the class from this

- Bayes' Theorem states that the **posterior** $P[\omega_i|x]$ is equal to the product of the **likelihood** $P[x|\omega_i]$ and the **prior** $P[\omega_i]$ divided by the **evidence** $P[x]$, or

$$P[\omega_i|x] = \frac{P[x|\omega_i]P[\omega_i]}{P[x]}$$

- Note, how this formula reverses the role of class ω_i and feature x

Bayes' Theorem

- The **posterior** is the probability of each class given a particular observed feature

$$P[\omega_i|x] = \frac{P[x|\omega_i]P[\omega_i]}{P[x]}$$

- Note, that the denominator is simply the sum over all classes of the numerator

$$P[x] = \sum_i P[x|\omega_i]P[\omega_i]$$

- Also note, that it is independent of the class and therefore equal for all classes; if it is not required to normalise the function to 1 (for example because we are only interested in the class with maximum posterior probability) then it is often omitted

Risk

- The **posterior** is the probability of each class given a particular observed feature

$$P[\omega_i|x] = \frac{P[x|\omega_i]P[\omega_i]}{P[x]}$$

- If we are taking an action α_j based on the assumption that we assume that class to be ω_i we use a **loss function**

$$L[\alpha_j, \omega_i]$$

- To define the **risk** of taking action α_j given observation x as

$$R[\alpha_j|x] = \sum_i L[\alpha_j|\omega_i]\,P[\omega_i|x]$$

Risk

- The conditional **risk** is the expected loss incurred by an action given and observed feature

$$R[\alpha_j|x] = \sum_i L[\alpha_j|\omega_i]\, P[\omega_i|x]$$

- In a Bayesian model the optimal decision rule $\alpha[x]$ is defined as minimising the overall risk

$$R = \int_x R[\alpha[x]|x]p[x]dx$$

- This can obviously be achieved by always selecting the action $\alpha_j \in \{\alpha_1, \dots, \alpha_k\}$ that minimises the conditional risk for every observed feature x

Risk

- The conditional **risk** is the expected loss incurred by an action given and observed feature

$$R[\alpha_j|x] = \sum_i L[\alpha_j|\omega_i]\, P[\omega_i|x]$$

- We already saw the 0/1-loss function in the context of classification errors

$$\mathrm{L}[\alpha_j, \omega_i] = \begin{cases} 0 & if\ j = i \\ 1 & if\ j \neq i \end{cases}$$

- The associated risk in this case depends only on the posterior of the associated class

$$R[\alpha_i|x] = 1 - P[\omega_i|x]$$

- Minimising this quantity is the same as maximising the posterior $P[\omega_i|x]$ leading to **minimum-error-rate classification**

Minimum-error-rate classification

- In summary: to classify a feature x such that the symmetric error-rate is minimised we have to maximising the posterior

$$P[\omega_i|x] = \frac{P[x|\omega_i]P[\omega_i]}{P[x]}$$

- Deciding on class ω_1 over ω_2 to be the more likely we look at

$$\frac{P[x|\omega_1]P[\omega_1]}{P[x]} > \frac{P[x|\omega_2]P[\omega_2]}{P[x]}$$

- which is equivalent to the **likelihood-ratio** being

$$\frac{P[x|\omega_1]}{P[x|\omega_2]} > \frac{P[\omega_2]}{P[\omega_1]}$$

Minimum-error-rate classification

- The optimal Bayesian decision rule for minimum-error-rate classification is to threshold the **likelihood-ratio** as follows

$$\frac{P[x|\omega_1]}{P[x|\omega_2]} > \frac{P[\omega_2]}{P[\omega_1]}$$

- For numerical reasons it is common to use logarithms of probabilities, so you will sometime find the equivalent formula

$$\log P[x|\omega_1] - \log P[x|\omega_2] > \log P[\omega_2] - logP[\omega_1]$$

Priors

- The optimal Bayesian decision rule for minimum-error-rate classification is to threshold the **likelihood-ratio** as follows

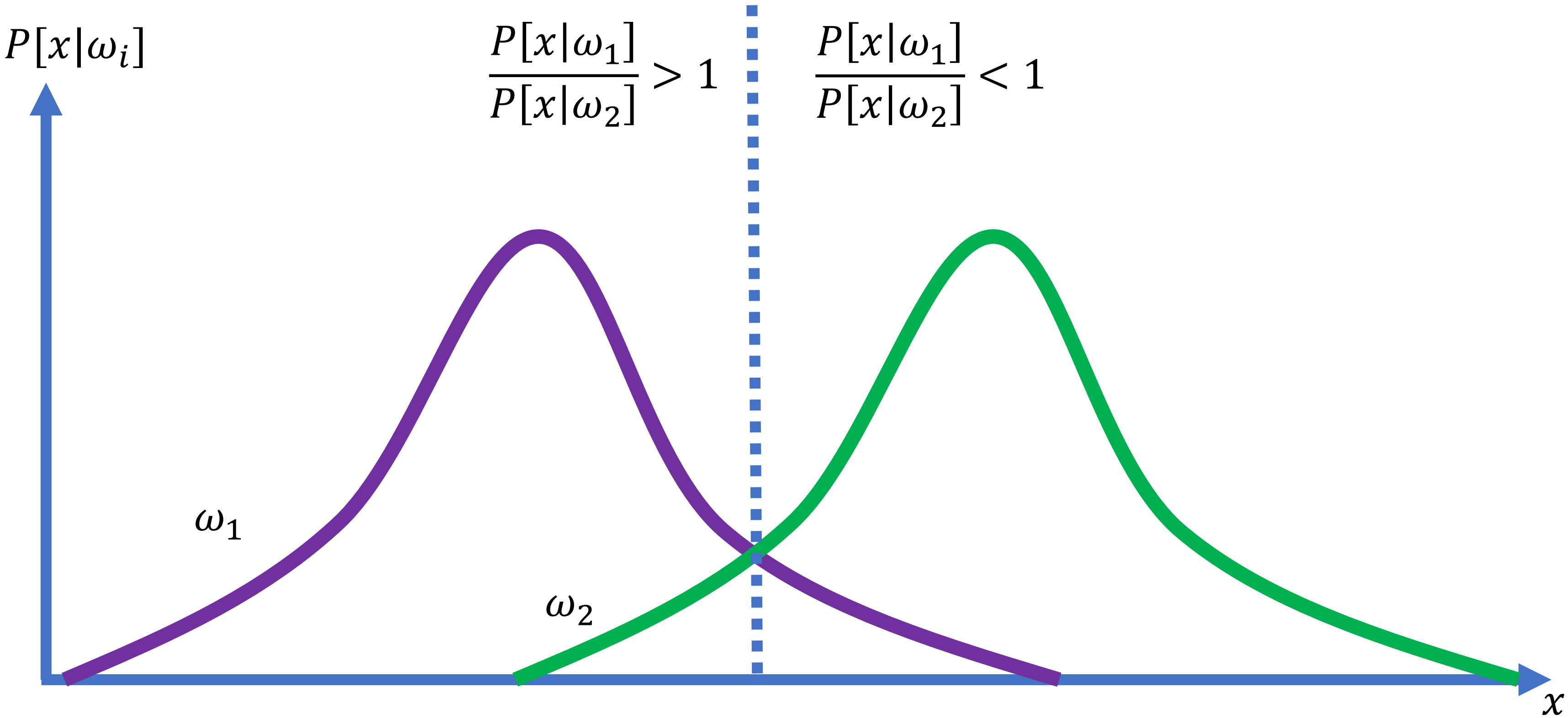
$$\frac{P[x|\omega_1]}{P[x|\omega_2]} > \frac{P[\omega_2]}{P[\omega_1]}$$

- The threshold is determined by the ratio of priors
- The **prior** $P[\omega_i]$ is the a-priori probability that you will encounter a specific class ω_i at all
- In case of random sampling the prior is determined by the relative class sizes, which can simply be counted
- If for example all classes have equal size (e.g. 50% male, 50% female), the threshold ratio is

$$\frac{P[\omega_2]}{P[\omega_1]} = 1$$

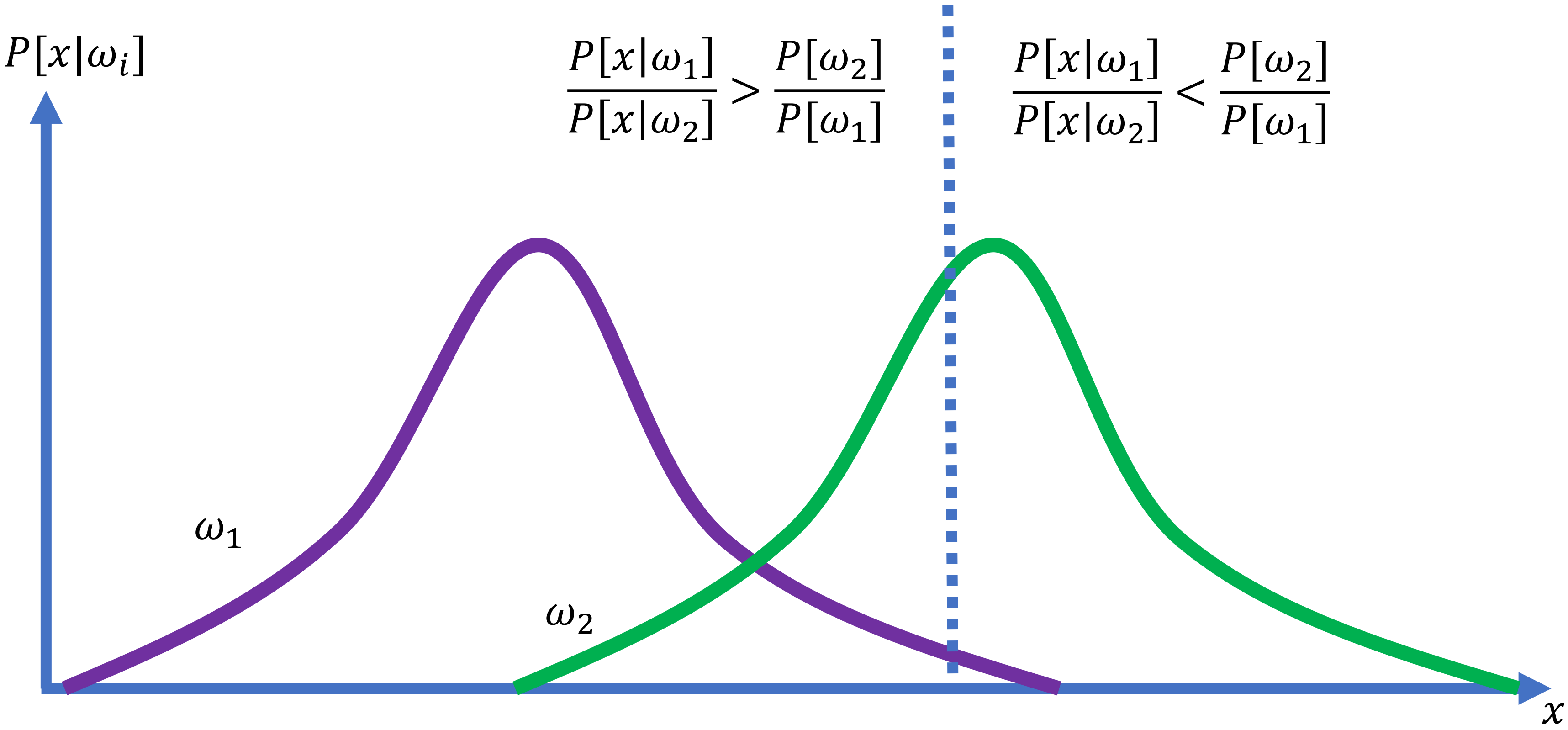
Bayesian decision rule

- In case of equal priors the Bayesian decision rule separates the feature space into areas depending on which likelihood is higher



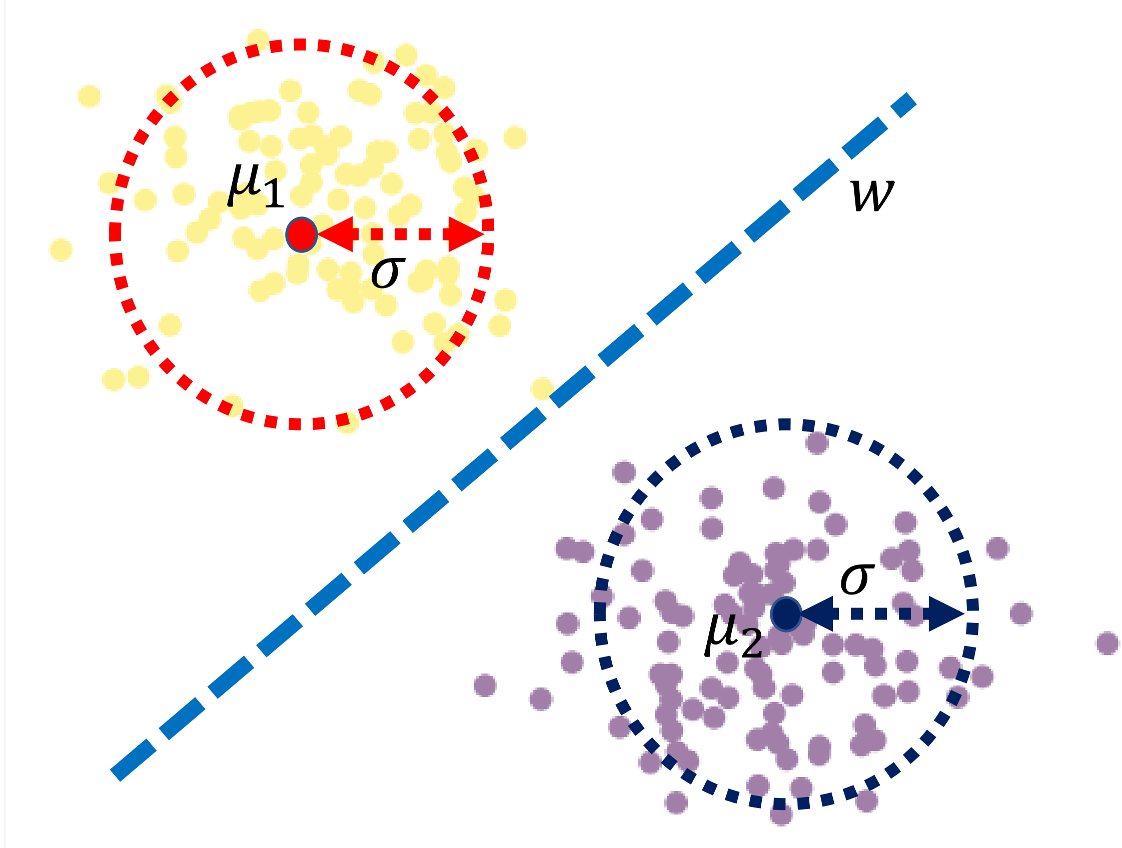
Bayesian decision rule

- The priors move this decision boundary towards the class that is a-priori less likely (ω_2), thereby favouring the more likely class (ω_1)



Example: n-d Normal densities

- Let’s assume we have two classes, and the likelihoods is given by two Normal densities with means μ_1 and μ_2 with equal and circular co-variances $\sigma^2 I$
- What is the optimal decision surface w



Example: n-d Normal densities

- The likelihood of the two distributions looks like

$$P[x|\lambda_i, \sigma^2] = \frac{\exp\left[-\frac{1}{2\sigma^2}(x-\mu_i)^T(x-\mu_i)\right]}{\sqrt{(2\pi)^n\sigma^2}}$$

- The likelihood-ratio is therefore

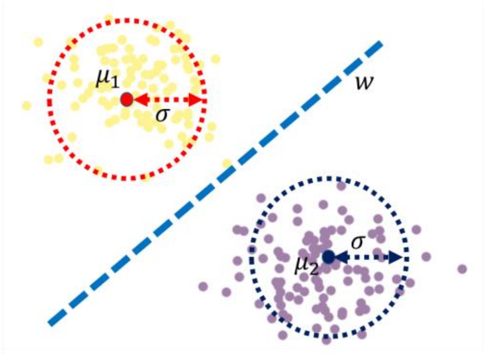
$$\frac{P[x|\lambda_1, \sigma^2]}{P[x|\lambda_2, \sigma^2]} = \exp\left[-\frac{1}{2\sigma^2}(2(\mu_2-\mu_1)^T x + \mu_1^T\mu_1 + \mu_2^T\mu_2)\right] > 1$$

- Taking logarithms yields the equivalent linear inequality

$$\underbrace{2(\mu_2-\mu_1)^T}_{w^T} x + \underbrace{\mu_1^T\mu_1 + \mu_2^T\mu_2}_{w_0} > 0$$

- showing that the optimal decision surface is the plane

$$w^T x + w_0 = 0$$



Thank you for your attention