Lecture 4

Bayesian Decision Theory

- Likelihood Ratio Test
- Gaussian distributions and the Likelihood Ratio Test
- Probability of error
- Bayes' risk

Perils of Over-fitting

Cross-Validation

Likelihood Ratio Test

Want to classify an object based on the evidence provided by a measurement (a feature vector) $\mathbf{x}$.

One decision rule is: Choose the class that is most probable given $\mathbf{x}$.

Mathematically this equates to choose class i such that

$$P(\omega_i|\mathbf{x}) \geq P(\omega_j|\mathbf{x}) \quad \text{for } j = 1, \ldots, C$$

For the 2-class problem the decision rule becomes:

$$\text{Class}\,(\mathbf{x}) = \begin{cases} \omega_1 & \text{if } P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x}) \\ \omega_2 & \text{if } P(\omega_1|\mathbf{x}) < P(\omega_2|\mathbf{x}) \end{cases}$$

Likelihood ratio test

This Bayesian decision rule can be re-written:

Choose class ω_1 if

$$P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x}),$$

$$\iff \quad \frac{p(\mathbf{x}\,|\,\omega_1)\,P(\omega_1)}{p(\mathbf{x})} > \frac{p(\mathbf{x}\,|\,\omega_2)\,P(\omega_2)}{p(\mathbf{x})}, \quad \text{Bayes' Rule}$$

$$\iff \quad p(\mathbf{x}\,|\,\omega_1)\,P(\omega_1) > p(\mathbf{x}\,|\,\omega_2)P(\omega_2), \quad \text{eliminate } p(\mathbf{x}) > 0$$

$$\iff \quad \underbrace{\frac{p(\mathbf{x}\,|\,\omega_1)}{p(\mathbf{x}\,|\,\omega_2)}}_{\text{likelihood ratio}} > \frac{P(\omega_2)}{P(\omega_1)}, \quad \text{as } P(\cdot) \geq 0$$

Introduce the notation

$$\Lambda(\mathbf{x}) \equiv \underbrace{\frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)}}_{\text{likelihood ratio}}$$

and the 2-class Bayesian decision rule / Likelihood Ratio Test can be written as

$$\text{Class}\,(\mathbf{x}) = \begin{cases} \omega_1 & \text{if } \Lambda(\mathbf{x}) > \frac{P(\omega_2)}{P(\omega_1)} \\ \omega_2 & \text{if } \Lambda(\mathbf{x}) < \frac{P(\omega_2)}{P(\omega_1)} \end{cases}$$

An example

Derive a **decision rule** for the 2-class problem based on the *Likelihood Ratio Test* assuming equal priors and class conditional densities:

$$p(x \,|\, \omega_1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-4)^2}{2}\right), \quad p(x \,|\, \omega_2) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-10)^2}{2}\right)$$

An example: Solution

Substitute the likelihoods and priors into the expressions in the LRT

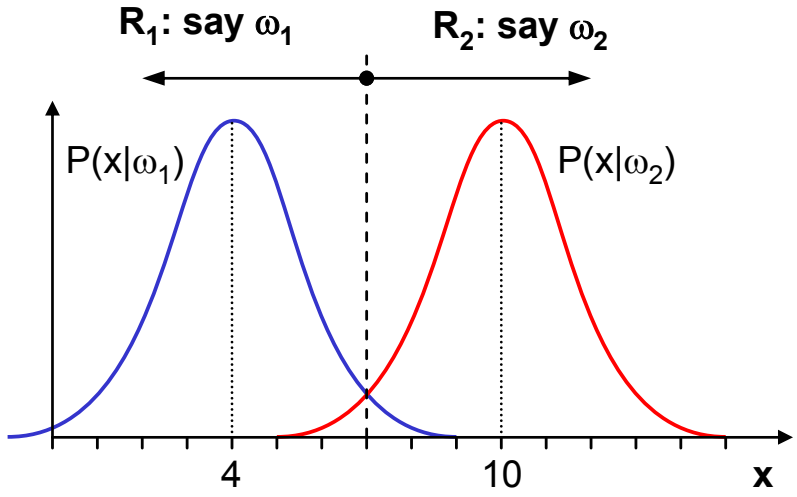
$$\Lambda(x) = \frac{(\sqrt{2\pi})^{-1} \exp\left(-.5(x-4)^2\right)}{(\sqrt{2\pi})^{-1} \exp\left(-.5(x-10)^2\right)}, \qquad \frac{P(\omega_2)}{P(\omega_1)} = \frac{.5}{.5} = 1$$

Choose class ω_1 if:

$$\begin{aligned}
& \Lambda(x) > 1 \\
\iff \quad & \exp\left(-.5(x-4)^2\right) > \exp\left(-.5(x-10)^2\right) \\
\iff \quad & (x-4)^2 < (x-10)^2, \quad \text{by taking logs and changing signs} \\
\iff \quad & x < 7
\end{aligned}$$

The LRT decision rule is:

$$\text{Class}\,(x) = \begin{cases} \omega_1 & \text{if } x < 7 \\ \omega_2 & \text{if } x > 7 \end{cases}$$



Question:

How does the LRT decision rule change if $P(\omega_1) = 2\,P(\omega_2)$?

Multi-variate example

Assume we have a 2 class problem but this time $\mathbf{x}$ is multi-variate and each $p(\mathbf{x} \mid \omega_i)$ is a multi-variate Gaussian:

$$p(\mathbf{x} \mid \omega_i) = \frac{1}{(2\,\pi)^{\frac{d}{2}}\,|\Sigma_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\,\Sigma_i^{-1}\,(\mathbf{x}-\boldsymbol{\mu}_i)\right)$$

Let the prior probability for each class be $P(\omega_1)$ and $P(\omega_2)$.

The likelihood ratio test says we choose class ω_1 when

$$\frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)} > \frac{P(\omega_2)}{P(\omega_1)}$$

$$\iff \frac{(2\,\pi)^{-\frac{d}{2}}\,|\Sigma_1|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\,\Sigma_1^{-1}\,(\mathbf{x}-\boldsymbol{\mu}_1)\right)}{(2\,\pi)^{-\frac{d}{2}}\,|\Sigma_2|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T\,\Sigma_2^{-1}\,(\mathbf{x}-\boldsymbol{\mu}_2)\right)} > \frac{P(\omega_2)}{P(\omega_1)}$$

$$\iff -\frac{1}{2}\log(|\Sigma_1|) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\,\Sigma_1^{-1}\,(\mathbf{x}-\boldsymbol{\mu}_1) \qquad \text{taking logs on both sides}$$

$$+\frac{1}{2}\log(|\Sigma_2|) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T\,\Sigma_2^{-1}\,(\mathbf{x}-\boldsymbol{\mu}_2)$$

$$> \log(P(\omega_2)) - \log(P(\omega_1))$$

$$\iff \frac{1}{2}\mathbf{x}^T\underbrace{(\Sigma_2^{-1}-\Sigma_1^{-1})}_{W}\mathbf{x} \;+\; \underbrace{(\boldsymbol{\mu}_1^T\Sigma_1^{-1}-\boldsymbol{\mu}_2^T\Sigma_2^{-1})}_{\mathbf{b}^T}\mathbf{x} \;+ \qquad \text{rearrange and group terms}$$

$$\underbrace{\frac{1}{2}\left(-\log(|\Sigma_1|)+\log(|\Sigma_2|)-\boldsymbol{\mu}_1^T\Sigma_1^{-1}\boldsymbol{\mu}_1+\boldsymbol{\mu}_2^T\Sigma_2^{-1}\boldsymbol{\mu}_2\right)-\log(P(\omega_2))+\log(P(\omega_1))}_{a}$$

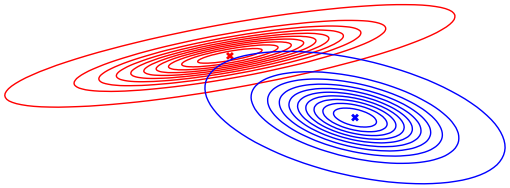
$$> 0$$

$$\equiv \frac{1}{2}\mathbf{x}^T W\mathbf{x} + \mathbf{b}^T\mathbf{x} + a > 0 \quad \Leftarrow \text{quadratic expression}$$

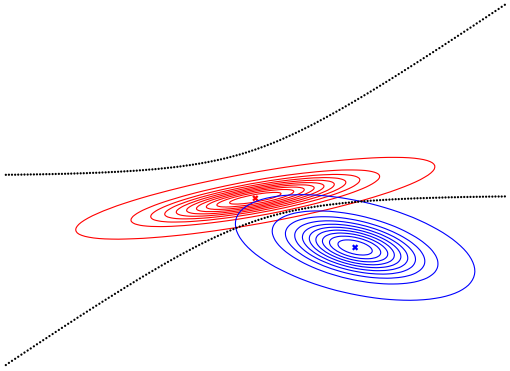
Bivariate example

Have a two class problem with

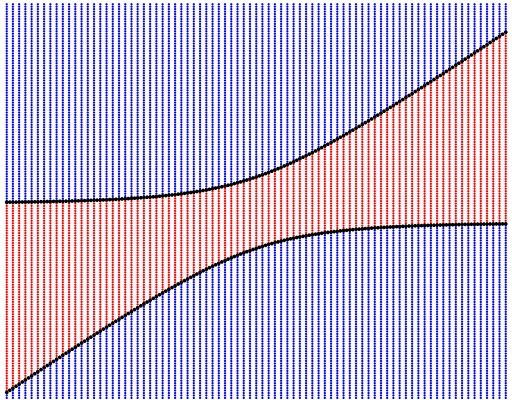
$$\boldsymbol{\mu}_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} .9 & .4 \\ .4 & .3 \end{pmatrix}, P(\omega_1) = .5 \qquad \boldsymbol{\mu}_2 = \begin{pmatrix} 2.6 \\ 3 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} .4 & -.2 \\ -.2 & .5 \end{pmatrix}, P(\omega_2) = .5$$



class distributions



decision boundaries

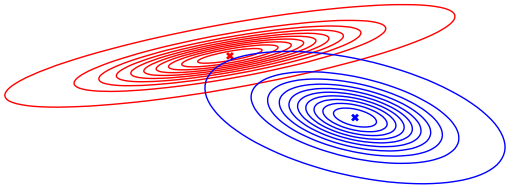


partition of space

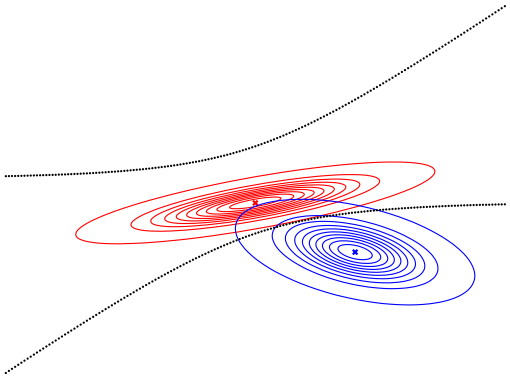
Bivariate example

Covariances same as the previous example but change in prior values

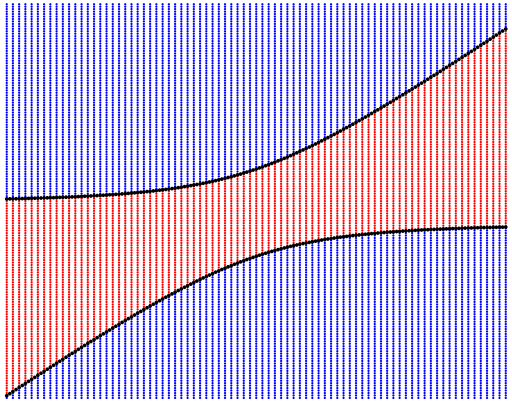
$$\boldsymbol{\mu}_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} .9 & .4 \\ .4 & .3 \end{pmatrix}, P(\omega_1) = .95 \qquad \boldsymbol{\mu}_2 = \begin{pmatrix} 2.6 \\ 3 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} .4 & -.2 \\ -.2 & .5 \end{pmatrix}, P(\omega_2) = .05$$



class distributions



decision boundaries

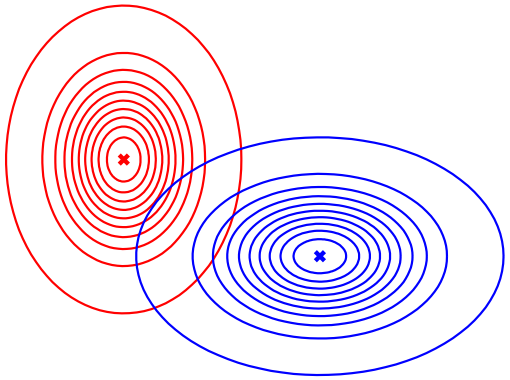


partition of space

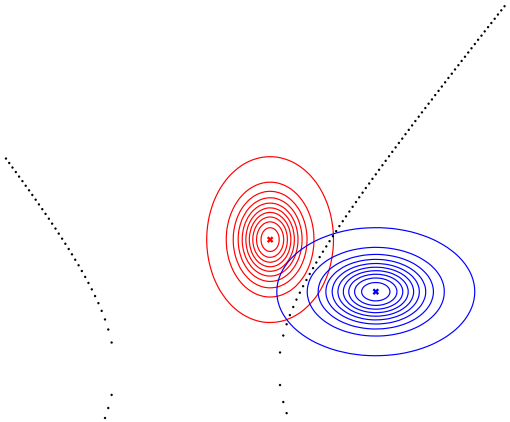
Another example

Have axis aligned covariance matrices:

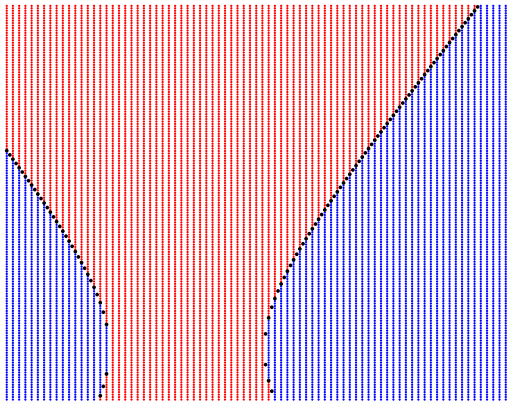
$$\boldsymbol{\mu}_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} .1 & 0 \\ 0 & 1.1 \end{pmatrix}, P(\omega_1) = .5 \qquad \boldsymbol{\mu}_2 = \begin{pmatrix} 2.6 \\ 3 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} .2438 & 0 \\ 0 & .6562 \end{pmatrix}, P(\omega_2) = .5$$



class distributions



decision boundaries

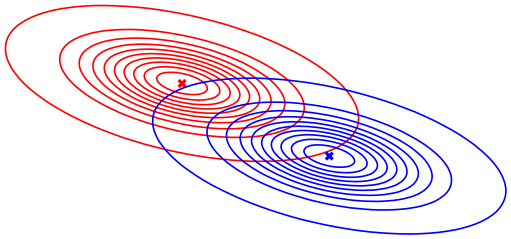


partition of space

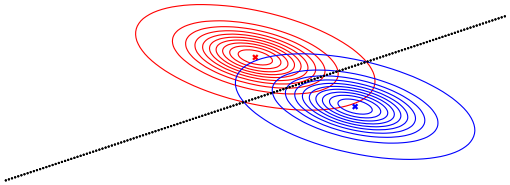
Another example

Each class has the same covariance matrix:

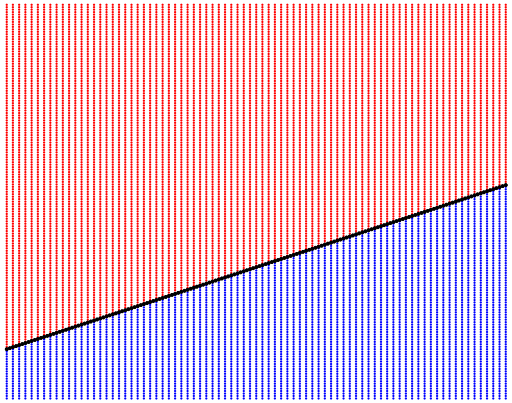
$$\boldsymbol{\mu}_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} .4 & -.2 \\ -.2 & .5 \end{pmatrix}, P(\omega_1) = .5 \qquad \boldsymbol{\mu}_2 = \begin{pmatrix} 2.6 \\ 3 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} .4 & -.2 \\ -.2 & .5 \end{pmatrix}, P(\omega_2) = .5$$



class distributions



decision boundaries



partition of space

What is the different with the decision boundary in this case?

Equal covariance matrices

Each class has the same covariance matrix

$\Longrightarrow$ decision boundary is a plane (line).

Remember, choose class ω_1 if $\boxed{\frac{1}{2}\mathbf{x}^T W\mathbf{x} + \mathbf{b}^T\mathbf{x} + a > 0}$ where

$$W = \Sigma_2^{-1} - \Sigma_1^{-1}$$
$$\mathbf{b}^T = (\boldsymbol{\mu}_1^T\Sigma_1^{-1} - \boldsymbol{\mu}_2^T\Sigma_2^{-1})$$
$$a = \ldots$$

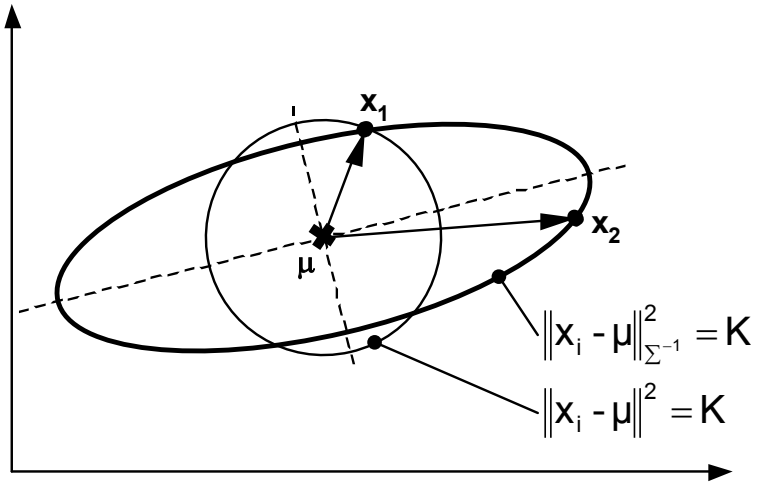
In the case of equal covariance matrices what is W equal to?

Write down the expression for the decision boundary.

An aside: Mahalanobis distance

You may have heard of the Mahalanobis distance. It is defined as

$$\|\mathbf{x}-\mathbf{y}\|^2_{\Sigma^{-1}} = (\mathbf{x}-\mathbf{y})^T\Sigma^{-1}(\mathbf{x}-\mathbf{y})$$



This is a vector distance that uses a Σ^{-1} norm.

- Σ^{-1} can be thought as a stretching factor on the space.
- For $\Sigma = I$ the Mahalanobis distance becomes the Euclidean distance.

Discriminant functions

All the decision rules presented in this lecture have the same structure:

> for each $\mathbf{x}$ in feature space choose class ω_i which maximizes (or minimizes) some measure $g_i(\mathbf{x})$

There is a set of discriminant functions $\{g_i(\mathbf{x})\}_{i=1}^{C}$ and decision rule

$$\text{assign } \mathbf{x} \text{ to class } \omega_i \text{ if } g_i(\mathbf{x}) > g_j(\mathbf{x}) \ \forall j \neq i$$

The three basic decision rules in this lecture are *Bayes*, *MAP* and *M*aximum*L*ikelihood in terms of *Discriminant Functions*:

Criterion	Discriminant Function
Bayes	$g_i(\mathbf{x}) = R(\alpha_i \mid \mathbf{x})$
MAP	$g_i(\mathbf{x}) = P(\omega_i \mid \mathbf{x})$
ML	$g_i(\mathbf{x}) = p(\mathbf{x} \mid \omega_i)$

Discriminant fns for Normally distributed classes

The MAP decision rule can be formulated as a family of discriminant functions

$$\text{Choose class } \omega_i \text{ if } g_i(\mathbf{x}) > g_j(\mathbf{x}) \ \forall i \neq j \text{ with } g_i(\mathbf{x}) = P(\omega_i \,|\, \mathbf{x})$$

For classes that are normally distributed, this family can be reduced to very simple expressions.

Using Bayes' rule the MAP discriminant function becomes

$$\begin{aligned} g_i(\mathbf{x}) &= \frac{p(\mathbf{x} \,|\, \omega_i)\, P(\omega_i)}{p(\mathbf{x})} \\ &= \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\mathbf{x}-\boldsymbol{\mu}_i)\right) \frac{P(\omega_i)}{p(\mathbf{x})} \end{aligned}$$

Eliminating constant terms

$$g_i(\mathbf{x}) = |\Sigma_i|^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) P(\omega_i)$$

Taking the $\log$ since it is a monotonically increasing function

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{1}{2}\log\left(|\Sigma_i|\right) + \log\left(P(\omega_i)\right)$$

How good is this decision rule?

Performance of a decision rule is measured by its

Probability of error:

$$P(\text{error}) = \sum_{i=1}^{C} P(\text{error} \,|\, \omega_i)\, P(\omega_i)$$

The class conditional probability of error is:

$$P(\text{error} \,|\, \omega_i) = \sum_{j \neq i} P(\text{choose } \omega_j \,|\, \omega_i) = \sum_{j \neq i} \int_{\mathcal{R}_j} p(\mathbf{x} \,|\, \omega_i)\, d\mathbf{x}$$

where $\mathcal{R}_j = \{\mathbf{x} : \text{Class}\,(\mathbf{x}) = \omega_j\}$.

For the 2-class problem

$$P(\text{error}) = P(\omega_1) \underbrace{\int_{\mathcal{R}_2} p(\mathbf{x} \,|\, \omega_1)\, d\mathbf{x}}_{\epsilon_1} + P(\omega_2) \underbrace{\int_{\mathcal{R}_1} p(\mathbf{x} \,|\, \omega_2)\, d\mathbf{x}}_{\epsilon_2}$$

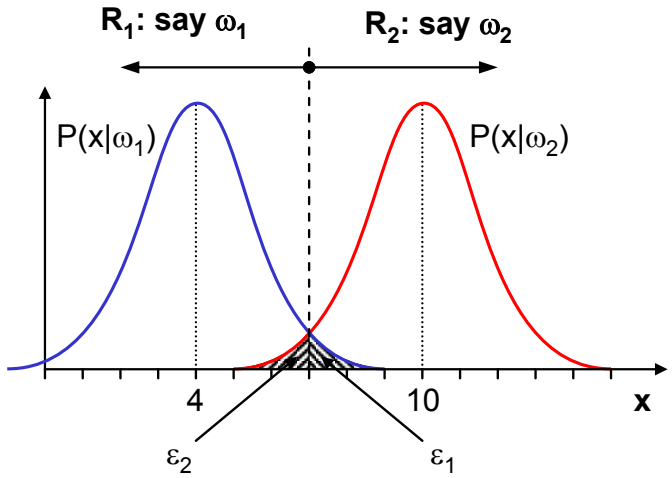
ϵ_1 is the integral of the likelihood $p(\mathbf{x} \,|\, \omega_1)$ over the region where ω_2 is chosen by the decision rule.

ϵ_2 is the integral of the likelihood $p(\mathbf{x} \,|\, \omega_2)$ over the region where ω_1 is chosen by the decision rule.

Back to the 1d example

For the decision rule of the previous example, the value of the integrals, ϵ_1 and ϵ_2, are depicted below.

Since we assumed equal priors, then $P(\text{error}) = .5\,(\epsilon_1 + \epsilon_2)$



Write out the expression for $P(\text{error})$ for this example.

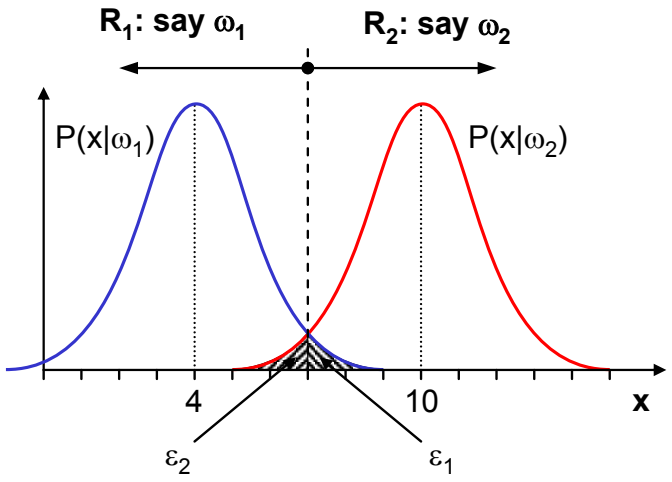
Back to the 1d example

The integrals ϵ_1 and ϵ_2 are:

$$\epsilon_1 = (2\pi)^{-\frac{1}{2}} \int_{x=7}^{\infty} \exp(-.5(x-4)^2)\, dx, \qquad \epsilon_2 = (2\pi)^{-\frac{1}{2}} \int_{x=-\infty}^{7} \exp(-.5(x-10)^2)\, dx$$

$$= .5 * (1 - \mathrm{erf}((7-4)/\sqrt{2})) \qquad = .5 + .5 * \mathrm{erf}((7-10)\sqrt{2}))$$

$$= .0013 \qquad = .0013$$



Probability of error

Thinking about the 2-class problem:

not all decisions are equally good wrt minimizing $P(\text{error})$.

For our example consider this (silly) rule:

$$\text{Class}\,(x) = \begin{cases} \omega_1 & \text{if } x < -100 \\ \omega_2 & \text{if } x \geq -100 \end{cases}$$

What is the $P(\text{error})$ for this decision rule?

Probability of error

Thinking about the 2-class problem:

not all decisions are equally good wrt minimizing $P(\text{error})$.

For our example consider this (silly) rule:

$$\text{Class}\,(x) = \begin{cases} \omega_1 & \text{if } x < -100 \\ \omega_2 & \text{if } x \geq -100 \end{cases}$$

For this rule $\epsilon_1 \approx 1$ and $\epsilon_2 \approx 0 \implies P(\text{error}) \approx .5$

The $P(\text{error})$ for the rule defined by the *likelihood ratio test* is .0013.

Thus the *likelihood ratio test* classifier is much better then our silly classifier in terms of minimizing the probability of error. In fact....

Bayes' Error Rate

For any decision problem, the minimum probability of error is achieved by the Likelihood Ratio Test decision rule. This probability of error is called the Bayes Error Rate and is the ***BEST*** any classifier can achieve.

Let's think for a moment

Our goal:

We are interested classifying and recognising images in a robust and efficient manner.

Recap:

You've just been told that once you've decided on the representation of your object then using a Bayes' classifier will minimize your probability of error.

Any problem here:

When can we achieve the Bayes Error Rate? What do we need to know explicitly? How often do we know these quantities exactly? Can we estimate the Bayes Error Rate from training data?

Estimating class conditional densities

Given labelled training data, from this will estimate the class conditional densities and build a classifier. If you know how you are going to model $p(\mathbf{x} \mid \omega_i)$ you must also

- know how to fit the parameters of my model from the training examples
- estimate $P(\omega_1)$ and $P(\omega_2)$

However, it is not always obvious what the best model is and how to recognize it. Therefore one must also

- estimate how well the classifier will perform on unseen data
- estimate the best value of the tunable parameters of this model

The learning problem

Hypothesis Class We consider some $restricted$ set $\mathcal{P}$ of probability functions $p : \mathcal{R}^d \rightarrow \mathcal{R}^+$ which in turn defines a set of mappings $f_p : \mathcal{R}^d \rightarrow \{0, 1\}$ via Bayes' Rule.

Estimation On the basis of a training set of examples and labels $\mathcal{X} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, find an estimate $\hat{p} \in \mathcal{P}$ and in turn $\hat{f} \equiv f_{\hat{p}}$.

Evaluation Measure how well $\hat{f}$ $generalizes$ to unseen examples, that is see whether $\hat{f}(\mathbf{x}_{\text{new}}) = y_{\text{new}}$ for a large number of $\mathbf{x}_{\text{new}}$

The problem of pdf parameter estimation is generally solved via maximum likelihood estimation (will go through this in more detail later on).

Training and test performance

Assume each training **and** test example-label pair, $(\mathbf{x}, y)$, is drawn *independently at random* from the *same* but unknown population of examples and labels. Represent this population as a joint pdf $p(\mathbf{x}, y)$.

Each example is a *sample* from this distribution $(\mathbf{x}_i, y_i) \sim p$. Then define

Empirical error (a.k.a. Training Error):

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f_{\hat{p}}(\mathbf{x}_i\,;\,\mathcal{X}))^2$$

Expected loss (a.k.a. Test Loss):

$$E_{(\mathbf{x},y)\sim p}\{L(\ y, f_{\hat{p}}(\mathbf{x}_i; \mathcal{X})\ )\}$$

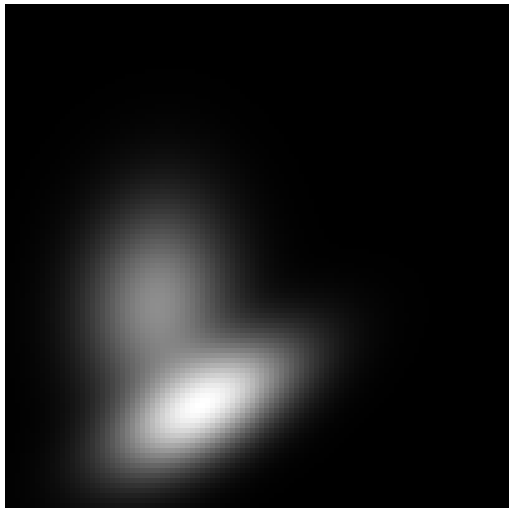
where $L(\ y, f_{\hat{p}}(\mathbf{x}_i; \mathcal{X})\ )$ is loss function which has a high value if the true label of an example does not match its predicted label and zero otherwise.

The training error based on a few sampled examples and labels serves as a proxy for the test performance measured over the whole population.

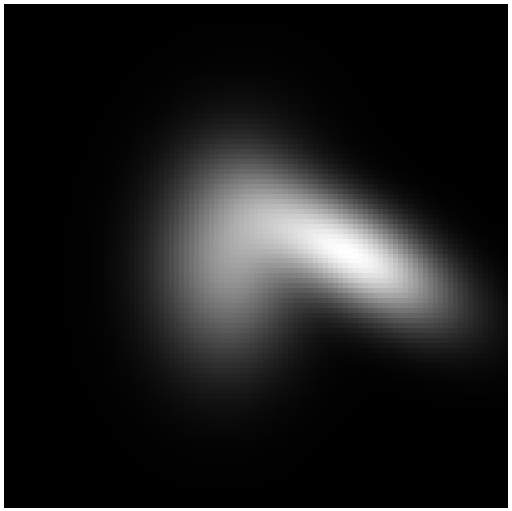
Is this a good idea?

Example

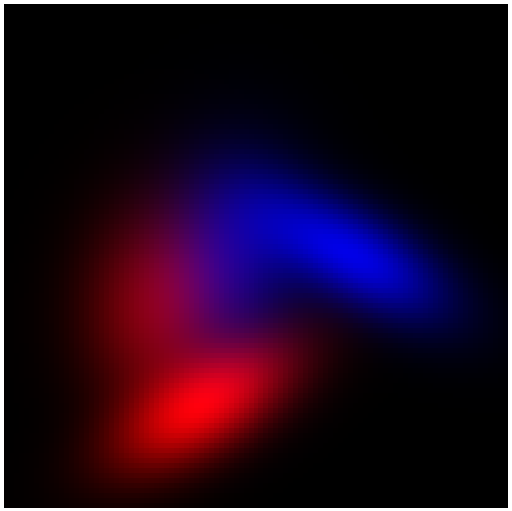
Consider these two class probability distributions which describe exactly the distribution of the feature vector for each class.



true $p(\mathbf{x} \mid \omega_1)$



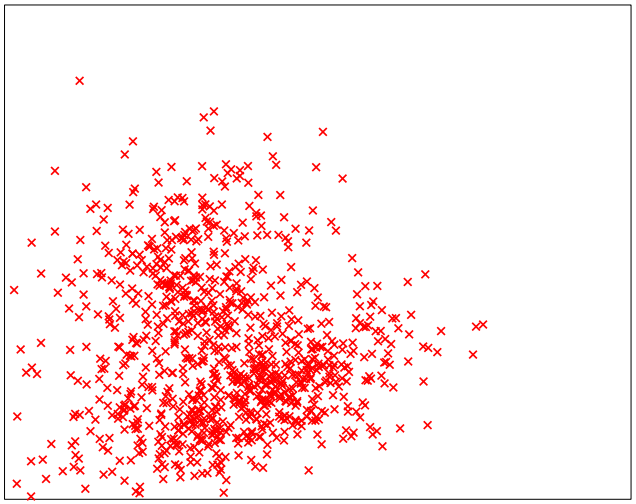
true $p(\mathbf{x} \mid \omega_2)$



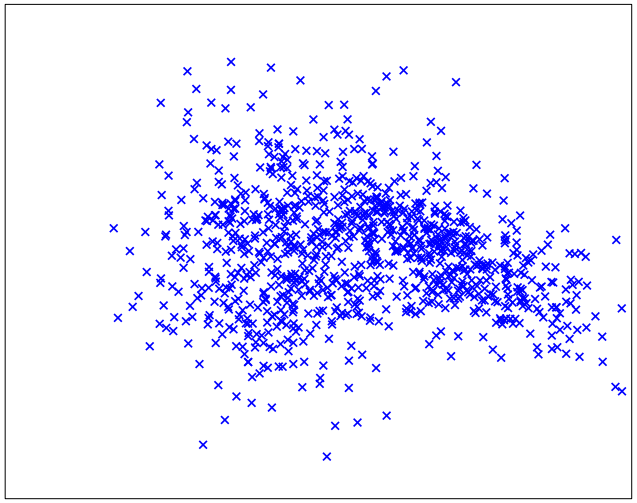
$p(\mathbf{x} \mid \omega_1)$ **and** $p(\mathbf{x} \mid \omega_2)$

Training Data

Initially have 1000 labelled training examples from each class. From this data we will estimate the $p(\mathbf{x} \mid \omega_i)$'s using 2d histograms:



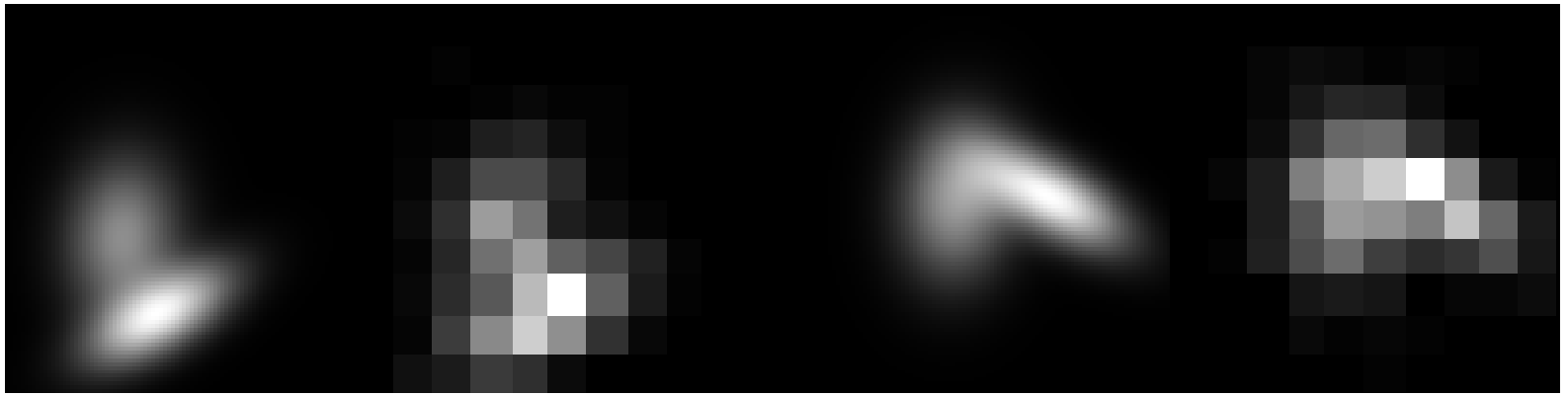
Class ω_1



Class ω_2

Estimate $p(\mathbf{x} \mid \omega_i)$'s

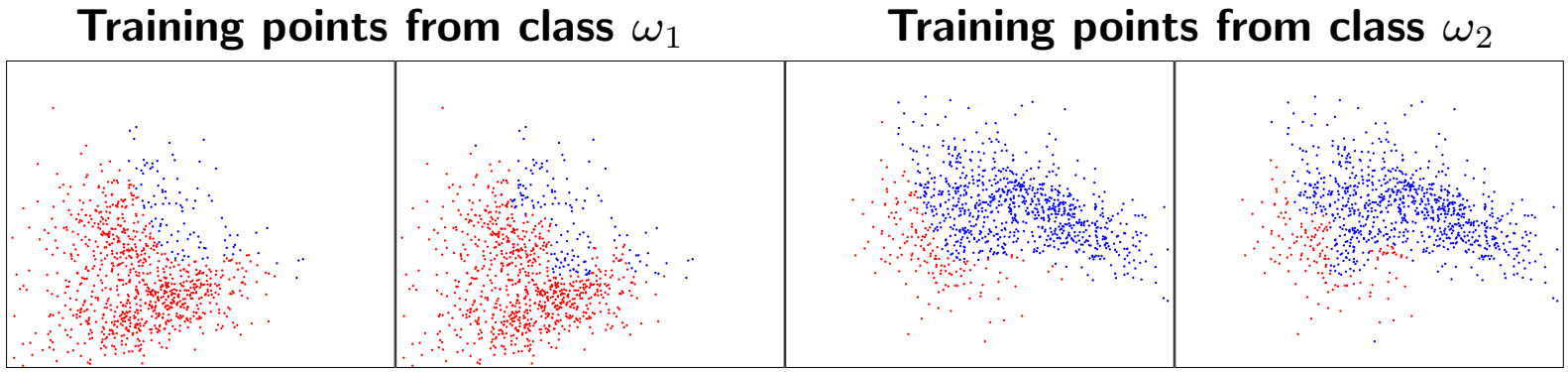
Calculate histograms from 1000 training points from each class. These estimate the class conditional probability distributions. Here we have a bin width of $.1$ in each dimension. This bin size defines which class of pdfs can be accurately model.



true $p(\mathbf{x} \mid \omega_1)$ **estimated** $\hat{p}(\mathbf{x} \mid \omega_1)$ **true** $p(\mathbf{x} \mid \omega_2)$ **estimated** $\hat{p}(\mathbf{x} \mid \omega_2)$

Classification results on training data

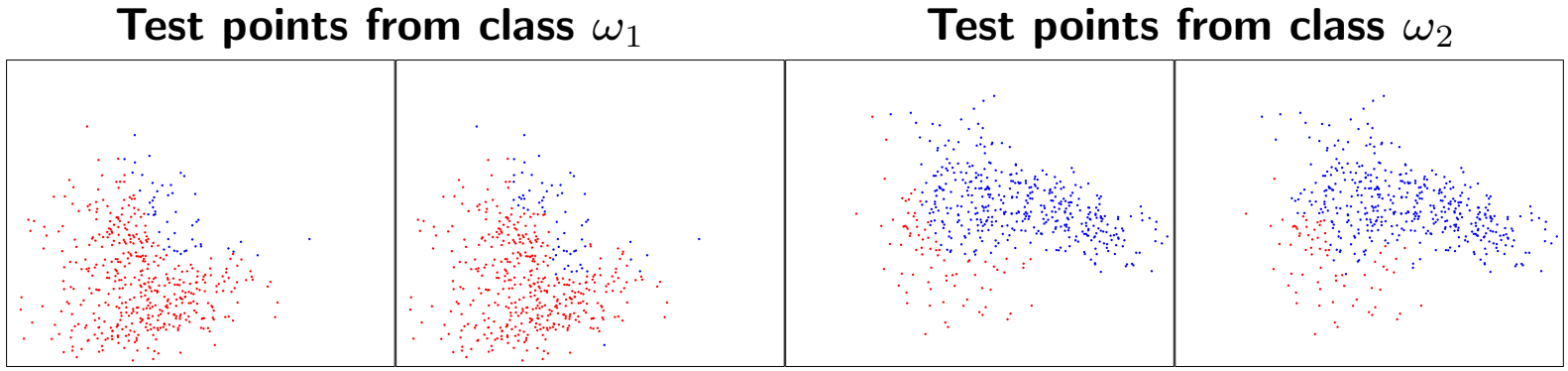
Classify, using the likelihood ratio test and the estimated $p(\mathbf{x} \mid \omega_i)$'s, the training points:



On this data the performance is okay. It is similar to that of the true Bayes' classifier.

Classification results

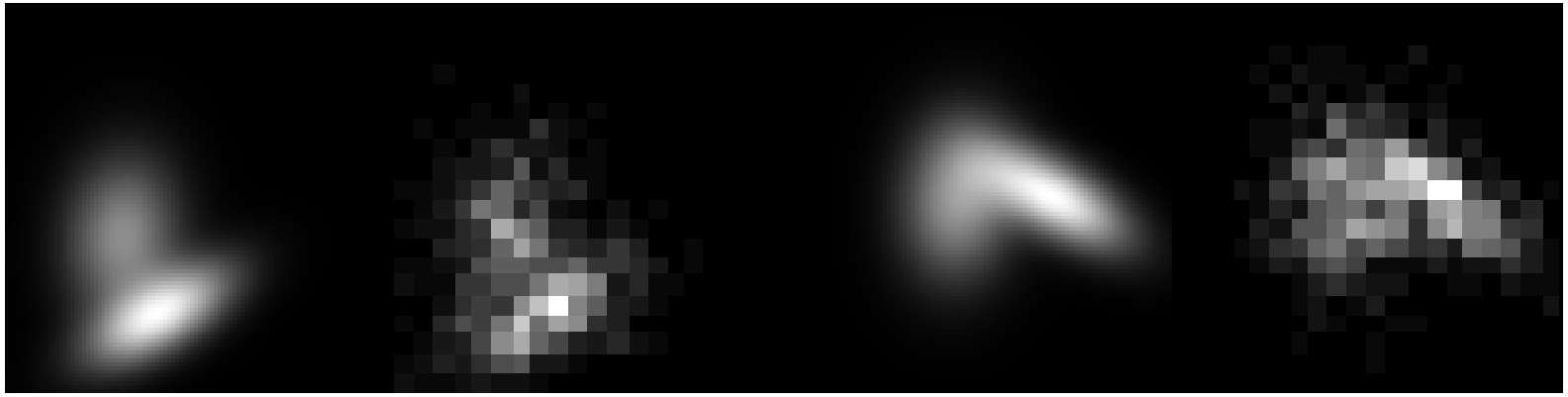
Classify, using the likelihood ratio test and the estimated $\hat{p}(\mathbf{x} \mid \omega_i)$'s, unseen test points generated by the true class conditionals:



Have good performance on the test data. Very similar to that of the true Bayes' classifier.

Estimate $p(\mathbf{x} \mid \omega_i)$'s

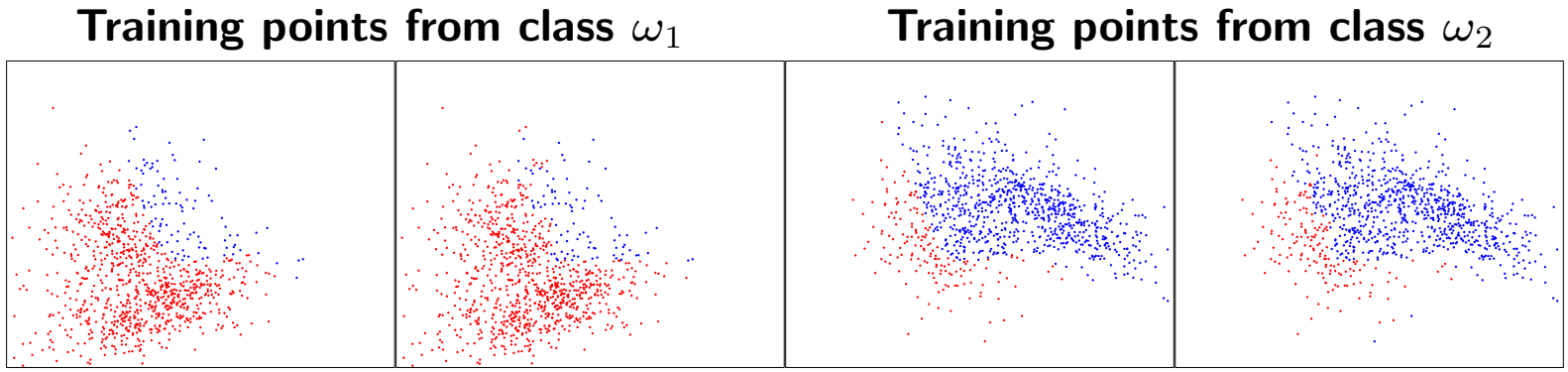
Calculate histograms from 1000 training points from each class. These estimate the class conditional probability distributions. Here we have a bin width of $.05$ in each dimension. Getting more detail..



true $p(\mathbf{x} \mid \omega_1)$ **estimated** $\hat{p}(\mathbf{x} \mid \omega_1)$ **true** $p(\mathbf{x} \mid \omega_2)$ **estimated** $\hat{p}(\mathbf{x} \mid \omega_2)$

Classification results on training data

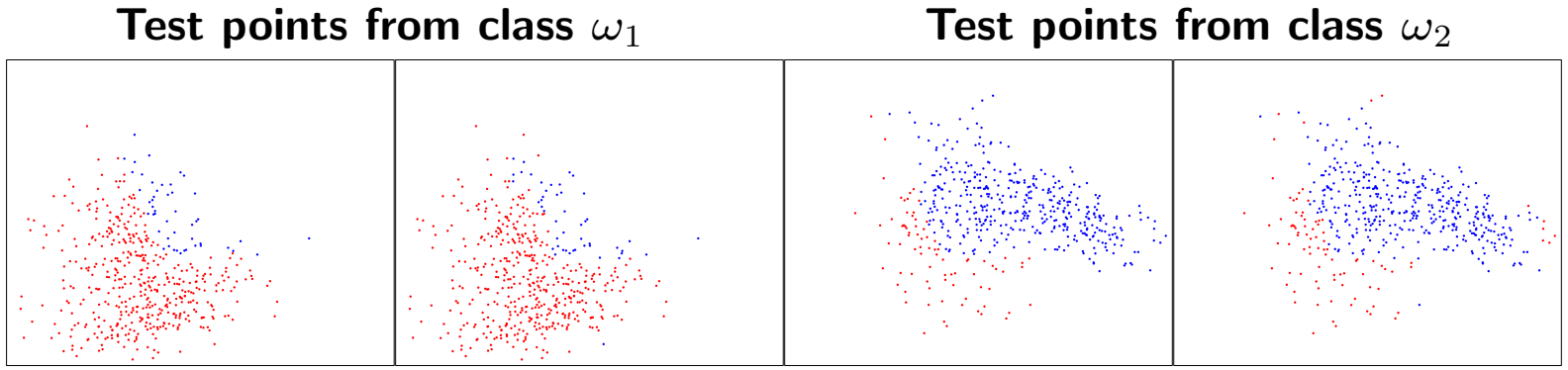
Classify, using the likelihood ratio test and the estimated $p(\mathbf{x} \mid \omega_i)$'s, the training points:



On this data the performance is okay. It is similar to that of the true Bayes' classifier.

Classification results

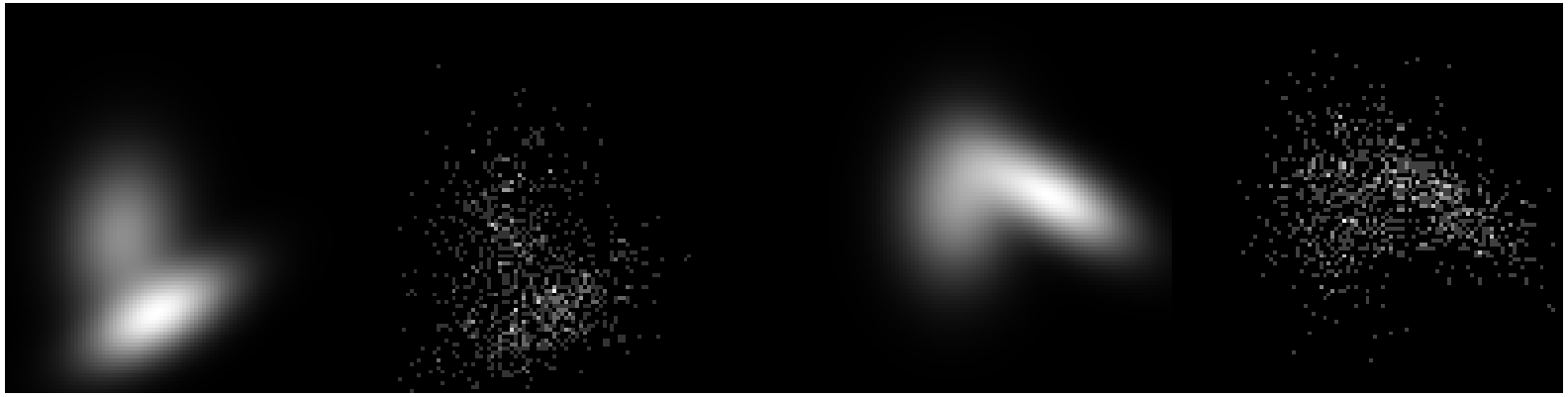
Classify, using the likelihood ratio test and the estimated $\hat{p}(\mathbf{x} \mid \omega_i)$'s, unseen test points generated by the true class conditionals:



Performance is quite satisfactory.

Estimate $p(\mathbf{x} \mid \omega_i)$'s

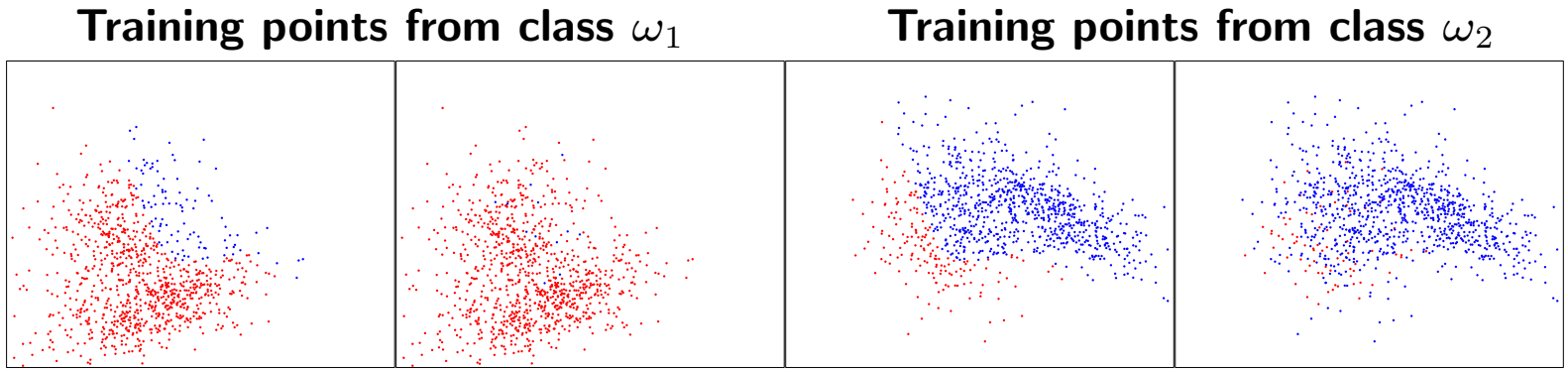
Calculate histograms from 1000 training points from each class. These estimate the class conditional probability distributions. Here we have a bin width of $.01$ in each dimension.



true $p(\mathbf{x} \mid \omega_1)$ **estimated** $\hat{p}(\mathbf{x} \mid \omega_1)$ **true** $p(\mathbf{x} \mid \omega_2)$ **estimated** $\hat{p}(\mathbf{x} \mid \omega_2)$

Classification results on training data

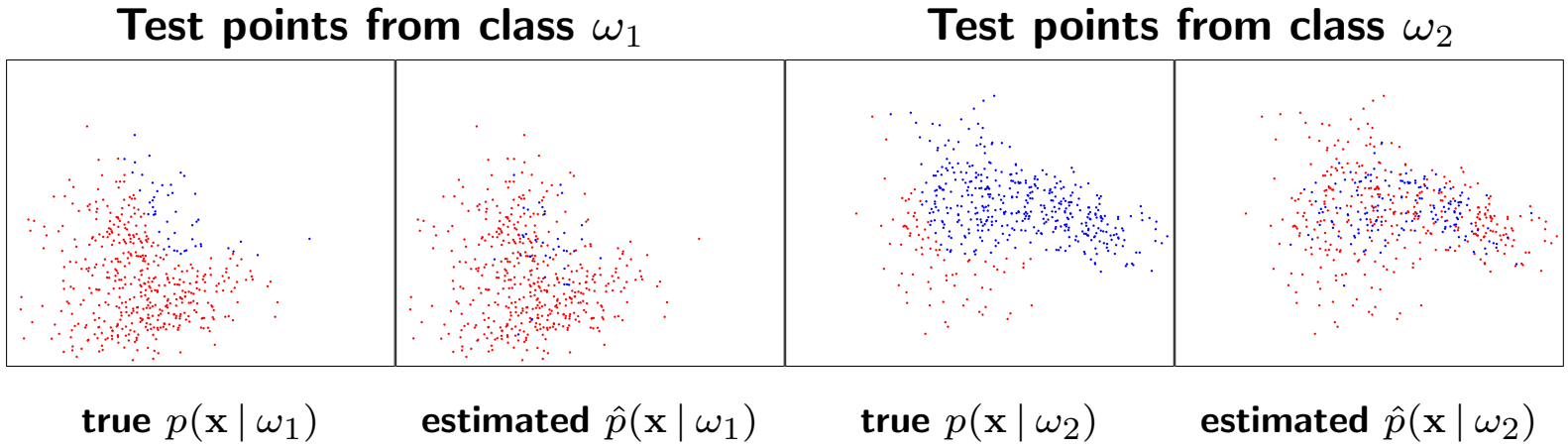
Classify, using the likelihood ratio test and the estimated $\hat{p}(\mathbf{x} \mid \omega_i)$'s, the training points:



Fantastic performance - should we be worried?

Classification results

Classify, using the likelihood ratio test and the estimated $\hat{p}(\mathbf{x} \mid \omega_i)$'s, unseen test points generated by the true class conditionals:



Terrible performance - we have **overfit** to the training data.

Complexity and over-fitting

With limited training examples our last histogram combined with Bayes' rule achieves classification with close to zero training error, but it has a large test (generalization) error.

$$\textbf{training error:} \quad \frac{1}{n}\sum_{t=1}^{n}(y_t - f_{\hat{p}}(\mathbf{x}_i;\mathcal{X}))^2 \approx 0$$

$$\textbf{test error:} \quad E_{(x,y)\sim P}(y - f_{\hat{p}}(\mathbf{x}_i;\mathcal{X}))^2 \gg 0$$

where $f_{\hat{p}}(\mathbf{x}_i;\mathcal{X})$ in this case is our estimated Bayes' classifier estimated from all the training data, $\mathcal{X}$.

Over-fitting occurs when

training error no longer bears any relation to the generalization error.

Warning: If your model is too flexible you will most likely over-fit.

Avoid over-fitting: cross validation

Cross Validation allows us to estimate the generalization error based on training examples alone.

Leave-one-out cross-validation treats each training example in turn as a test example:

$$CV = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_{\hat{p}}(\mathbf{x}_i \;\; \mathcal{X}\backslash(\mathbf{x}_i, y_i)))^2$$

where $f_{\hat{p}}(\mathbf{x}_i;\; \mathcal{X}\backslash(\mathbf{x}_i, y_i))$ in this case is our estimated Bayes' classifier estimated from all the training data $\mathcal{X}$ except data-point $(\mathbf{x}_i, y_i)$.

Avoid over-fitting: cross validation

Cross Validation allows us to estimate the generalization error based on training examples alone.

Leave-one-out cross-validation is quite computationally expensive so another cross validation technique is K-**fold cross-validation**. In this case the training data is partitioned into K sets - $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_K$ and

$$CV = \frac{1}{n} \sum_{i=1}^{K} \sum_{(\mathbf{x},y) \in \mathcal{X}_i} (y - f_{\hat{p}}(\mathbf{x}; \mathcal{X} \backslash \mathcal{X}_i))^2$$

where $f_{\hat{p}}(\mathbf{x}_i; \mathcal{X} \backslash \mathcal{X}_i)$, in this case, is our estimated Bayes' classifier estimated from all the training data $\mathcal{X}$ minus $\mathcal{X}_i$.

Cross validation: model selection

As cross validation gives an estimate of the generalization error of a classifier it can be used to estimate from a set of classifiers which one performs the best.

So in our histogram example cross-validation could be used to estimate the best bin width for the distribution we are trying to model.

Naive Bayes

For dimensions of $\mathbf{x} = (x_1, \ldots, x_d)$ greater than 3, we have a problem modelling it with a histogram as one is increasingly prone to over-fitting. Thus frequently one models all the dimensions as independent given the class ω_i

$$p(\mathbf{x} \,|\, \omega_i) = \prod_{j=1}^{d} p(x_j \,|\, \omega_i)$$

The posterior is expressed as

$$P(\omega_i \,|\, \mathbf{x}) \propto p(\mathbf{x} \,|\, \omega_i) P(\omega_i) = P(\omega_i) \prod_{j=1}^{d} p(x_j \,|\, \omega_i)$$

Building a classifier based on this posterior is known as performing Naive Bayes.

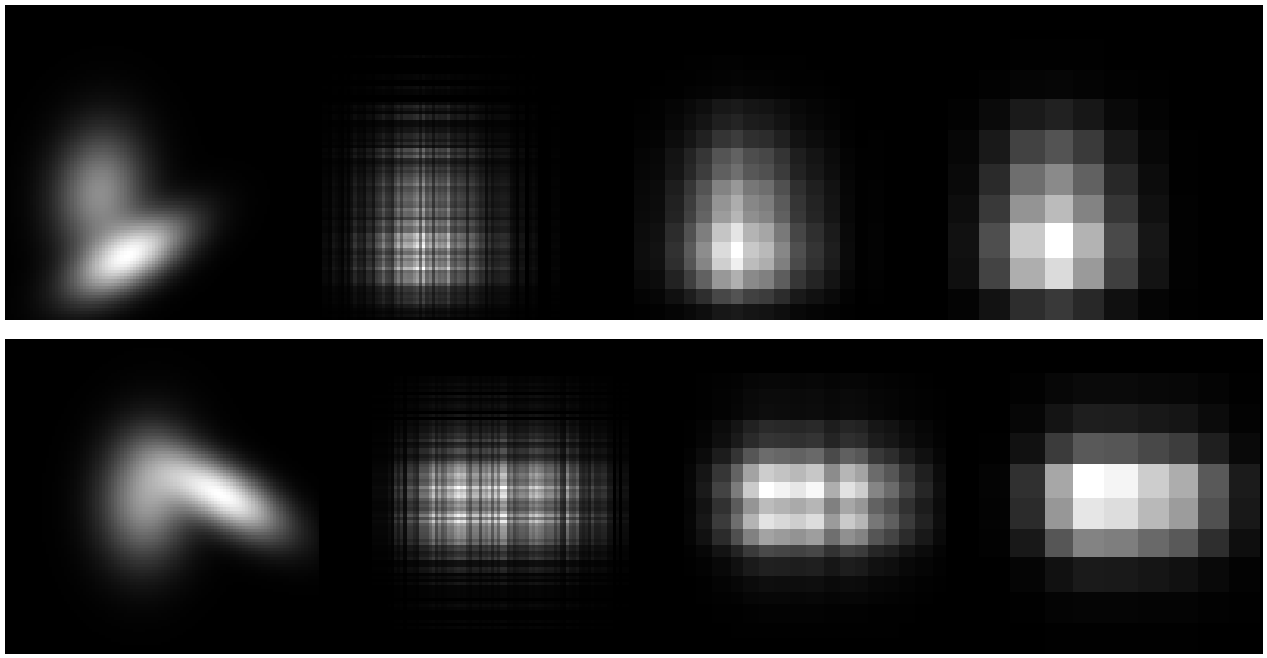
Naive Bayes is very computationally cheap and allows the use of very high dimensional $\mathbf{x}$. And in many cases it can produce surprisingly good classification results as it is not particularly prone to over-fitting.

If $p(\mathbf{x} \mid \omega_i) \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, Naive Bayes is equivalent to assuming what about Σ?

Given many training $\mathbf{x}_i$'s what ways have we learnt to estimate $p(\mathbf{x} \mid \omega_i)$?

Estimate $p(\mathbf{x} \mid \omega_i)$'s with Naive Bayes

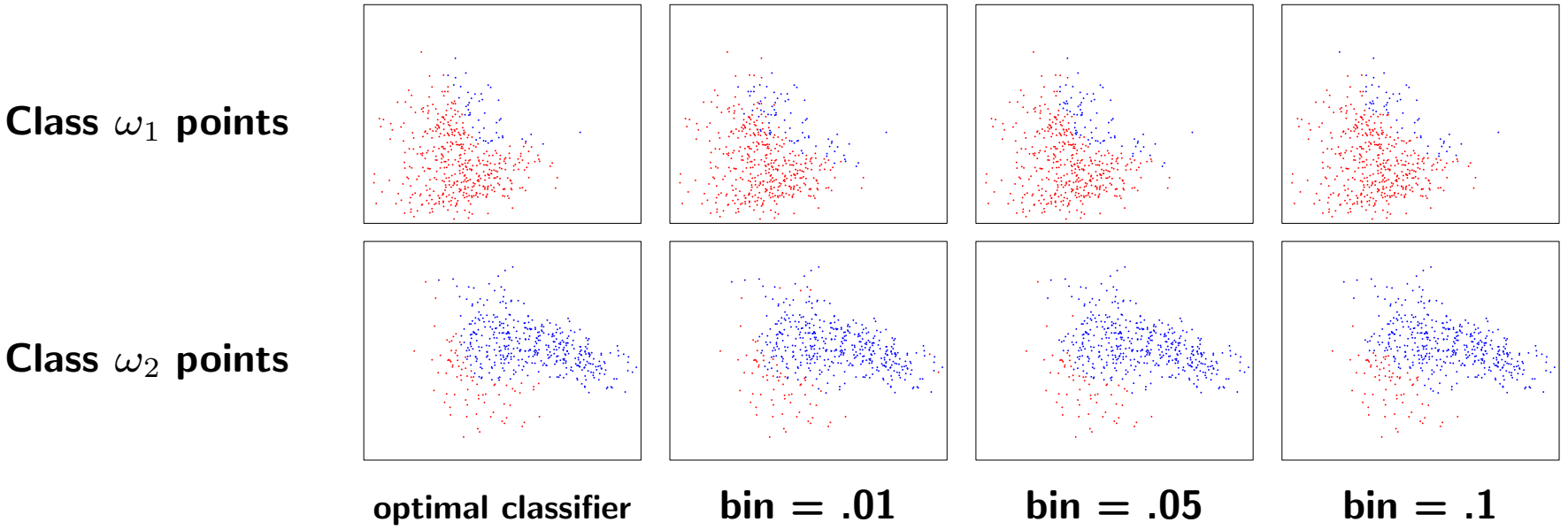
Return to our previous data and fit a histogram independently to each dimension and multiply them together for the joint likelihood.



true distributions **bin = .01** **bin = .05** **bin = .1**

Classification results with Naive Bayes

Classify unseen test points generated by the true class conditionals, using the likelihood ratio test and the naively estimated $p(\mathbf{x} \mid \omega_i)$'s.



Differences to the previous results??

Bayes' Risk

Which misclassification is worse and why ?

- classifying a faulty airplane as a safe airplane
- classifying a safe airplane as a faulty airplane

Bayes' Risk

Which misclassification is worse and why ?

- classifying a faulty airplane as a safe airplane
  puts people's lives in danger
- classifying a safe airplane as a faulty airplane
  costs the airline company money

Not all misclassifications are equal!

Bayes' Risk

Formalize this concept in terms of a cost function C_{ij}.

> Let C_{ij} denote the cost of choosing class ω_i when ω_j is the true class.

Bayes' Risk is the expected value of the cost

$$\mathrm{E}\left[C\right] = \sum_{i=1}^{2}\sum_{j=1}^{2} C_{ij}\, P(\text{decide } \omega_i,\ \omega_j \text{ true class})$$

$$= \sum_{i=1}^{2}\sum_{j=1}^{2} C_{ij}\, p(\mathbf{x} \in \mathcal{R}_i \,|\, \omega_j)\, P(\omega_j)$$

Bayes' Risk

What is the decision rule that minimizes the Bayes' Risk ?

- First note: $p(\mathbf{x} \in \mathcal{R}_i \,|\, \omega_j) = \int_{\mathbf{x} \in \mathcal{R}_i} p(\mathbf{x} \,|\, \omega_j)\, d\mathbf{x}$
- Bayes' Risk is equal to:

$$\mathrm{E}\,[C] = \int_{\mathcal{R}_1} [C_{11}\, P(\omega_1)\, p(\mathbf{x} \,|\, \omega_1) + C_{12}\, P(\omega_2)\, p(\mathbf{x} \,|\, \omega_2)]\, d\mathbf{x} + \int_{\mathcal{R}_2} [C_{21}\, P(\omega_1)\, p(\mathbf{x}|\omega_1) + C_{22}\, P(\omega_2)\, p(\mathbf{x} \,|\, \omega_2)]\, d\mathbf{x}$$

- Now remember

$$\int_{\mathcal{R}_1} p(\mathbf{x} \,|\, \omega_j)\, d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x} \,|\, \omega_j)\, d\mathbf{x} = \int_{\mathcal{R}_1 \cup \mathcal{R}_2} p(\mathbf{x} \,|\, \omega_j)\, d\mathbf{x} = 1.$$

$$
\begin{aligned}
\mathrm{E}\left[C\right] = {} & C_{11}\, P(\omega_1) \int_{\mathcal{R}_1} p(\mathbf{x} \mid \omega_1)\, d\mathbf{x} + C_{12}\, P(\omega_2) \int_{\mathcal{R}_1} p(\mathbf{x} \mid \omega_2)\, d\mathbf{x} + \\
& C_{21}\, P(\omega_1) \int_{\mathcal{R}_2} p(\mathbf{x} \mid \omega_1)\, d\mathbf{x} + C_{22}\, P(\omega_2) \int_{\mathcal{R}_2} p(\mathbf{x} \mid \omega_2)\, d\mathbf{x} + \\
& \boxed{C_{21}\, P(\omega_1) \int_{\mathcal{R}_1} p(\mathbf{x} \mid \omega_1)\, d\mathbf{x} + C_{22}\, P(\omega_2) \int_{\mathcal{R}_1} p(\mathbf{x} \mid \omega_2)\, d\mathbf{x}} + \leftarrow +A \\
& \boxed{-C_{21}\, P(\omega_1) \int_{\mathcal{R}_1} p(\mathbf{x} \mid \omega_1)\, d\mathbf{x} - C_{22}\, P(\omega_2) \int_{\mathcal{R}_1} p(\mathbf{x} \mid \omega_2)\, d\mathbf{x}} \leftarrow -A \\
= {} & C_{21}\, P(\omega_1) \int_{\mathcal{R}_1 \cup \mathcal{R}_2} p(\mathbf{x} \mid \omega_1)\, d\mathbf{x} + C_{22}\, P(\omega_2) \int_{\mathcal{R}_1 \cup \mathcal{R}_2} p(\mathbf{x} \mid \omega_2)\, d\mathbf{x} + \\
& (C_{12} - C_{22})\, P(\omega_2) \int_{\mathcal{R}_1} p(\mathbf{x} \mid \omega_2)\, d\mathbf{x} - (C_{21} - C_{11})\, P(\omega_1) \int_{\mathcal{R}_1} p(\mathbf{x} \mid \omega_1)\, d\mathbf{x} \\
= {} & C_{21}\, P(\omega_1) + C_{22}\, P(\omega_2) + \\
& (C_{12} - C_{22})\, P(\omega_2) \int_{\mathcal{R}_1} p(\mathbf{x} \mid \omega_2)\, d\mathbf{x} - (C_{21} - C_{11})\, P(\omega_1) \int_{\mathcal{R}_1} p(\mathbf{x} \mid \omega_1)\, d\mathbf{x}
\end{aligned}
$$

We want to find the region $\mathcal{R}_1$ that minimizes the Bayes' Risk. From the previous slide we see the first two terms of $\mathrm{E}[C]$ are constant with respect to $\mathcal{R}_1$. Thus the optimal region is:

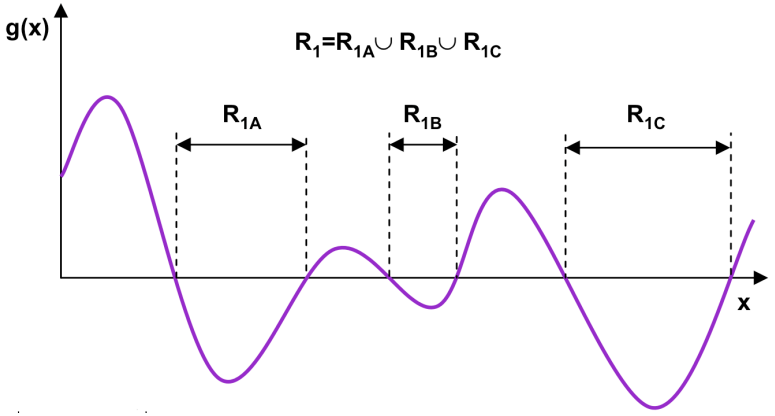
$$\begin{aligned}\mathcal{R}_1^* &= \arg\min_{\mathcal{R}_1} \left\{ \int_{\mathcal{R}_1} \left[(C_{12} - C_{22}) P(\omega_2) p(\mathbf{x} \mid \omega_2) - (C_{21} - C_{11}) P(\omega_1) p(\mathbf{x} \mid \omega_1) \right] d\mathbf{x} \right\} \\ &= \arg\min_{\mathcal{R}_1} \left\{ \int_{\mathcal{R}_1} g(\mathbf{x})\, d\mathbf{x} \right\}\end{aligned}$$

Note we are assuming $C_{21} > C_{11}$ and $C_{12} > C_{22}$, that is the cost of a misclassification is higher than the cost of a correct classification. Thus:

$$(C_{12} - C_{22}) > 0 \quad \text{AND} \quad (C_{21} - C_{11}) > 0$$

Bayes' Risk (2)

Momentarily forget about the specific expression of $g(\mathbf{x})$. Consider the type of decision region $\mathcal{R}_1^*$ we are looking for. The intervals that minimize the integral $\int_{\mathcal{R}_1} g(\mathbf{x})\, d\mathbf{x}$ are those where $g(\mathbf{x}) < 0$



Thus choose $\mathcal{R}_1^*$ such that

$$(C_{21} - C_{11})P(\omega_1)p(\mathbf{x}\,|\,\omega_1) > (C_{12} - C_{22})P(\omega_2)p(\mathbf{x}\,|\,\omega_2)$$

Rearranging the terms yields:

$$\frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)} > \frac{(C_{12} - C_{22})P(\omega_2)}{(C_{21} - C_{11})P(\omega_1)}$$

Therefore we obtain the decision rule:

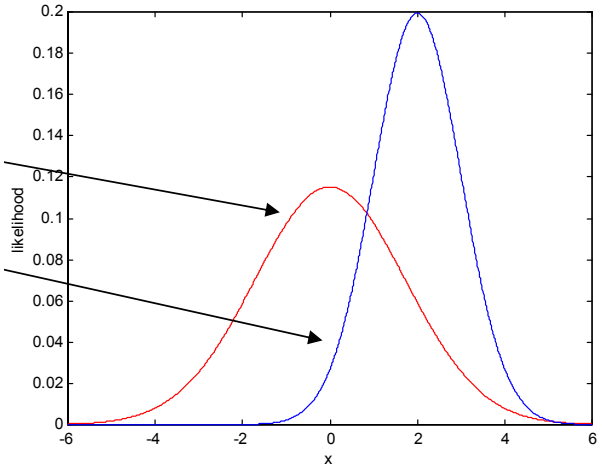
A Likelihood Ratio Test

$$\text{Class}\,(\mathbf{x}) = \begin{cases} \omega_1 & \text{if } \frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)} > \frac{(C_{12} - C_{22})P(\omega_2)}{(C_{21} - C_{11})P(\omega_1)} \\ \omega_2 & \text{if } \frac{p(\mathbf{x} \mid \omega_1)}{p(\mathbf{x} \mid \omega_2)} < \frac{(C_{12} - C_{22})P(\omega_2)}{(C_{21} - C_{11})P(\omega_1)} \end{cases}$$

Bayes' Risk: An example

Consider the following 2 class classification problem. The likelihood functions for each class are:

$$p(x \mid \omega_1) = (2\pi 3)^{-\frac{1}{2}} \exp\left(-.5\, x^2/3\right), \quad p(x \mid \omega_2) = (2\pi)^{-\frac{1}{2}} \exp\left(-.5\, (x-2)^2\right)$$



The priors are: $P(\omega_1) = P(\omega_2) = .5$

Define the (mis)classification costs as: $C_{11} = C_{22} = 0, C_{12} = 1, C_{21} = \sqrt{3}$

Problem: Determine a decision rule minimizing the Bayes' risk.

Bayes' Risk: an example

Solution: $$\Lambda(x) = \frac{(2\pi 3)^{-\frac{1}{2}} \exp\left(-.5x^2/3\right)}{(2\pi)^{-\frac{1}{2}} \exp\left(-.5(x-2)^2\right)} = \frac{(3)^{-\frac{1}{2}} \exp\left(-.5x^2/3\right)}{\exp\left(-.5(x-2)^2\right)}.$$

Choose class ω_1 if $\Lambda(x) > \dfrac{.5(1-0)}{.5(\sqrt{3}-0)}$

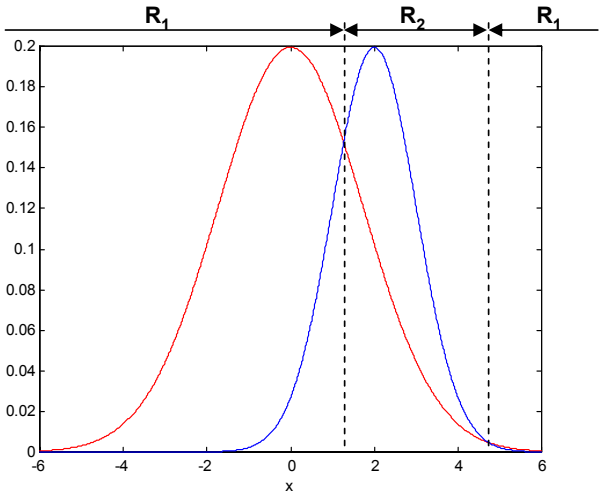
$$\iff \frac{(3)^{-\frac{1}{2}} \exp\left(-.5x^2/3\right)}{\exp\left(-.5(x-2)^2\right)} > 1$$

$$\iff \exp\left(-.5x^2/3\right) > \exp\left(-.5(x-2)^2\right)$$

$$\iff -\frac{1}{2}\frac{x^2}{3} > \frac{1}{2}(x-2)^2$$

$$\iff x^2 - 6x + 6 > 0$$

$$\iff x > 4.73 \text{ and } x < 1.27$$



Today's programming assignment

Programming Assignment

- Details available on the course website.
- You will write `Matlab` functions to fit a multi-variate Gaussian distribution to skin colour data and also to *background* data. Using these Gaussian models you will then classify unseen pixels as skin or non-skin based on a likelihood ratio test.
- Mail me about any errors you spot in the Exercise notes.
- I will notify the class about errors spotted and corrections via the course website and mailing list.