Decision Theory

Pattern recognition

Francesco Corona

Outline

Decision theory

Decision theory

Probability theory provides a consistent framework for quantifying and manipulating uncertainty

Decision theory combined with probability theory allows us to make optimal decisions

- in situations involving uncertainty

We have an input vector $\mathbf{x}$ together with a corresponding target vector $\mathbf{t}$

- our goal is to predict $\mathbf{t}$ given a new value for $\mathbf{x}$

The joint probability distribution $p(\mathbf{x}, \mathbf{t})$ provides a complete summary of the uncertainty associated with these variables

- its determination from a set of training data is an example of **inference**
- it is typically a very difficult problem, it is the main topic of PRML book

Decision theory (cont.)

- For regression problems, **t** will comprise continuous variables
- For classification problems, **t** will be representing class labels

We must **make a specific prediction** for the value of **t**, or more generally, **take a specific action** based on our understanding of the values **t** is likely to take

Decision theory (cont.)

Consider a medical diagnosis problem in which we have taken an X-ray image of a patient, and we wish to determine whether the patient has a disease or not

- The input vector $\mathbf{x}$ is given by the set of pixel intensities in the image
- The output variable t will represent the presence/absence of disease

We denote:

- the presence of disease by the class $\mathcal{C}_1$
- the absence of disease by the class $\mathcal{C}_2$

We also choose t to be a binary variable

- $t = 0$ corresponds to class $\mathcal{C}_1$
- $t = 1$ corresponds to class $\mathcal{C}_2$

Decision theory (cont.)

The general inference problem then involves determining the joint distribution $p(\mathbf{x}, \mathcal{C}_k)$, or $p(\mathbf{x}, t)$, which gives us the most complete probabilistic description

- This can be a very useful and informative quantity, cool

In the end, we must decide whether to give treatment to the patient or not

- We want to make **optimal decisions** in some proper sense
- The **decision step** is the central subject of decision theory
- How to make optimal decisions given **proper probabilities**

The decision is generally simple, provided that we have solved the inference

Decision theory (cont.)

When we obtain the image $\mathbf{x}$ for a new patient, our goal is to decide which of the two classes to assign to the image

- We are interested in the probabilities of the two classes, given the image
- Using Bayes' theorem, these probabilities $p(\mathcal{C}_k|\mathbf{x})$ can be expressed as

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} \tag{1}$$

$p(\mathcal{C}_k)$ is the prior probability for class $\mathcal{C}_k$ and $p(\mathcal{C}_k|\mathbf{x})$ its corresponding posterior

- $p(\mathcal{C}_1)$ is the probability that the patient has disease, before we even take the X-ray measurement
- $p(\mathcal{C}_1|\mathbf{x})$ is the corresponding probability, revised using Bayes in the light of the X-ray information

All quantities in Equation 1 can be obtained from the joint probability $p(\mathbf{x}, \mathcal{C}_k)$

Decision theory (cont.)

If our aim is to minimise the chance of assigning $\mathbf{x}$ to the wrong class, then intuitively we would choose the class having the highest posterior probability

- This intuition is clearly correct, but why?

Minimising the misclassification rate

Decision theory

Minimising the misclassification rate (cont.)

Suppose that our goal is simply to make as few misclassifications as possible

- Firstly, we need a rule that assigns $\mathbf{x}$ to one of the available classes

Such rule will divide the input space into **decision regions** $\mathcal{R}_k$

- one for each class $\mathcal{C}_k$ and such that points in $\mathcal{R}_k$ are assigned to class $\mathcal{C}_k$

The boundaries between regions are called **decision boundaries**

Mistakes when inputs $\mathbf{x}$ belonging to class $\mathcal{C}_1$ ($\mathcal{C}_2$) are assigned to class $\mathcal{C}_2$ ($\mathcal{C}_1$)

- The probability of this occurring is given by

$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) d\mathbf{x} \end{aligned} \tag{2}$$

Minimising the misclassification rate (cont.)

We are free to choose whatever decision rule that assigns $\mathbf{x}$ to either class

If we want to minimise $p(\text{mistake})$, we should make sure that $\mathbf{x}$ is assigned to whichever class has the smallest value of the integrand

$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) d\mathbf{x} \end{aligned}$$

- If $p(\mathbf{x}, \mathcal{C}_1) > p(\mathbf{x}, \mathcal{C}_2)$ for a given $\mathbf{x}$, then we should assign it to class $\mathcal{C}_1$
- If $p(\mathbf{x}, \mathcal{C}_2) > p(\mathbf{x}, \mathcal{C}_1)$ for a given $\mathbf{x}$, then we should assign it to class $\mathcal{C}_2$

Minimising the misclassification rate (cont.)

- If $p(\mathbf{x}, \mathcal{C}_1) > p(\mathbf{x}, \mathcal{C}_2)$ for a given $\mathbf{x}$, then we should assign it to class $\mathcal{C}_1$
- If $p(\mathbf{x}, \mathcal{C}_2) > p(\mathbf{x}, \mathcal{C}_1)$ for a given $\mathbf{x}$, then we should assign it to class $\mathcal{C}_2$

From the product rule of probability, we have that $p(\mathbf{x}, \mathcal{C}_k) = p(\mathcal{C}_k|\mathbf{x})p(\mathbf{x})$

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}, \mathcal{C}_k)}{p(\mathbf{x})}$$

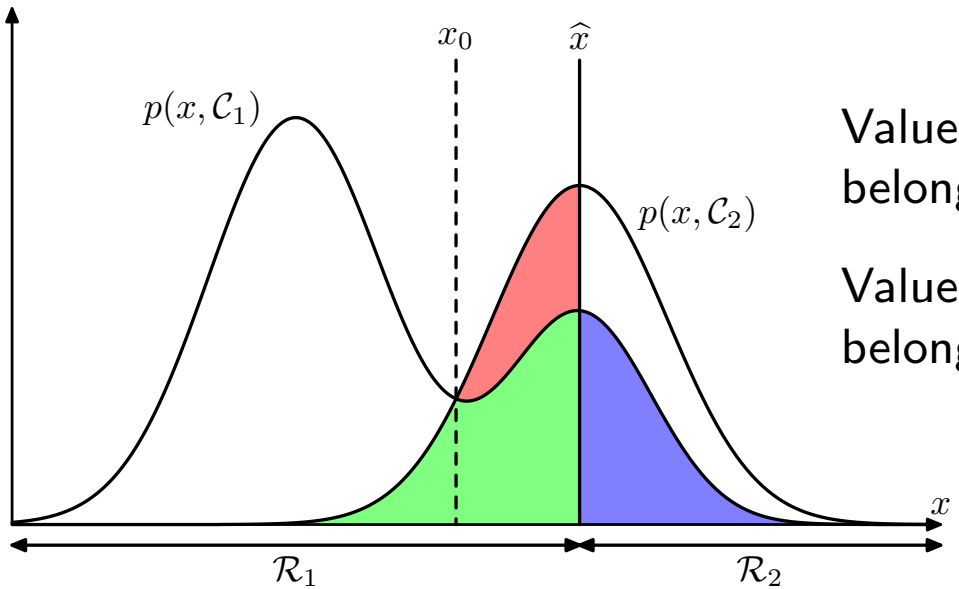
- If $p(\mathcal{C}_1|\mathbf{x}) > p(\mathcal{C}_2|\mathbf{x})$ for a given $\mathbf{x}$, then we should assign it to class $\mathcal{C}_1$
- If $p(\mathcal{C}_2|\mathbf{x}) > p(\mathcal{C}_1|\mathbf{x})$ for a given $\mathbf{x}$, then we should assign it to class $\mathcal{C}_2$

The mistake probability is minimum when each $\mathbf{x}$ is assigned to the class for which the posterior probability $p(\mathcal{C}_k|\mathbf{x})$ is largest

Minimising the misclassification rate (cont.)

The joint probabilities $p(x, \mathcal{C}_k)$ for each of two classes plotted against x

- together with the decision boundary $x = \hat{x}$



Values of $x \geq \hat{x}$ are classified as $\mathcal{C}_2$ and belong to decision region $\mathcal{R}_2$

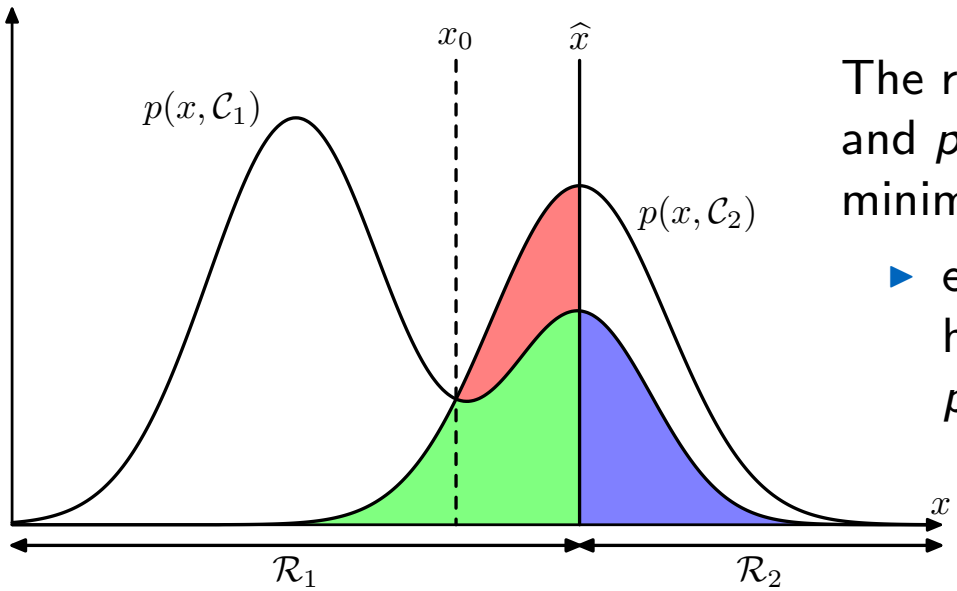
Values of $x < \hat{x}$ are classified as $\mathcal{C}_1$ and belong to decision region $\mathcal{R}_1$

Minimising the misclassification rate (cont.)

Errors arise from the blue, green, and red regions, so that

- for $x < \hat{x}$, errors from points from class $\mathcal{C}_2$ are misclassified as $\mathcal{C}_1$
  - the sum of red and green areas
- for $x \geq \hat{x}$, errors from points from class $\mathcal{C}_1$ are misclassified as $\mathcal{C}_2$
  - the blue area

As we move the decision boundary the combined blue and green area remains constant while the red area varies, the optimal $\hat{x}$ is when the red area disappears



The red area vanishes when $p(x, \mathcal{C}_1)$ and $p(x, \mathcal{C}_2)$ cross and we have a minimum misclassification rate

- each x is assigned to the class with highest posterior probability $p(\mathcal{C}_k|x)$

Minimising the misclassification rate (cont.)

With K classes it is easier to maximise the probability of being correct

$$p(\text{correct}) = \sum_{k=1}^{K} p(\mathbf{x} \in \mathcal{R}_k, \mathcal{C}_k) = \sum_{k=1}^{K} \int_{\mathcal{R}_k} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x} \tag{3}$$

It is maximised when the regions $\mathcal{R}_k$ are chosen such that each $\mathbf{x}$ is assigned to the class for which $p(\mathbf{x}, \mathcal{C}_k)$ is largest

Again, using the product rule $p(\mathbf{x}, \mathcal{C}_k) = p(\mathcal{C}_k|\mathbf{x})p(\mathbf{x})$ and noting that the factor of $p(\mathbf{x})$ is common to all terms

- each $\mathbf{x}$ should be assigned to the class with the largest posterior probability $p(\mathcal{C}_k|\mathbf{x})$

Minimising the expected loss

Decision theory

Minimising the expected loss

The objective can be more complex than simply minimising misclassifications

In the medical diagnosis example:

1. if a patient who does not have disease is incorrectly diagnosed as having it, the consequences may be patient distress plus further investigations
2. if a patient who does have disease is diagnosed as healthy, the result may be premature death due to lack of treatment

The consequences of these two types of mistake can be dramatically different

- It would clearly be better to make fewer mistakes of the second kind, even if this was at the expense of making more mistakes of the first kind

Minimising the expected loss (cont.)

We can formalise such issues through the introduction of a **loss function**

- A single, overall measure of loss incurred in taking any available decision

Our goal is then to minimise the total loss incurred [1]

Suppose that, for a new value of $\mathbf{x}$, the true class is $\mathcal{C}_k$ and that we assign $\mathbf{x}$ to class $\mathcal{C}_j$, where j may or may not be equal to k

- in doing so, we incur some level of loss, here L_{kj}
- L_{kj} can view as the k, j element of a **loss matrix**

$$\begin{array}{c} \\ \text{cancer} \\ \text{normal} \end{array} \begin{array}{c} \begin{array}{cc} \text{cancer} & \text{normal} \end{array} \\ \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix} \end{array}$$

No loss if the correct decision is made
Loss of 1 if a healthy patient is diagnosed as ill
Loss of 1000 if a ill patient is diagnosed as fine

[1] Note that some consider equivalently a **utility function**, whose value they aim to maximise

Minimising the expected loss (cont.)

The optimal solution is clearly now the one which minimises the loss function

- The loss function depends on the true class, which is unknown
- Given an input $\mathbf{x}$, our uncertainty in the true class is $p(\mathbf{x}, \mathcal{C}_k)$

We try to **minimise the average loss**, where the average is computed with respect to the joint probability distribution $p(\mathbf{x}, \mathcal{C}_k)$

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x} = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} \Big(p(\mathcal{C}_k|\mathbf{x}) p(\mathbf{x}) \Big) d\mathbf{x} \qquad (4)$$

- Each $\mathbf{x}$ can be assigned independently to one decision region $\mathcal{R}_j$
- Our goal is to choose regions that minimise the expected loss
- For each $\mathbf{x}$, minimise $\sum_k L_{kj} p(\mathbf{x}, \mathcal{C}_k)$ with $p(\mathbf{x}, \mathcal{C}_k) = p(\mathbf{x}|\mathcal{C}_k) p(\mathbf{x})$

Minimising the expected loss (cont.)

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) d\mathbf{x} = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} \Big(p(\mathcal{C}_k|\mathbf{x}) p(\mathbf{x}) \Big) d\mathbf{x}$$

After eliminating the common factor $p(\mathbf{x})$ resulting from the application of the product rule, the decision rule that minimises the expected loss is the one that assigns each new $\mathbf{x}$ to the class j from which the following quantity is minimum

$$\sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x}) \tag{5}$$

Which is simple to do, once we know the **posterior class probability** $p(\mathcal{C}_k|\mathbf{x})$

Reject option
Decision theory

Reject option

Classification errors arise from some specific regions of input space

- where the largest of the posteriors $p(\mathcal{C}_k|\mathbf{x})$ is very less than 1

or equivalently

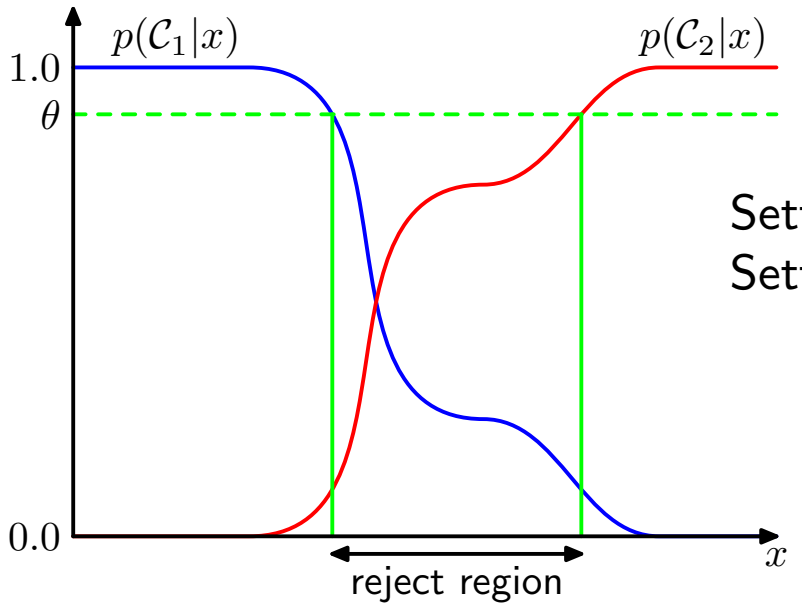
- where the joint distributions $p(\mathbf{x}, \mathcal{C}_k)$ have comparable values

In regions where we are relatively uncertain about class membership

- It is appropriate to avoid making decisions on such cases
- This is known as the **reject option**
- In difficult cases, we leave the human do the classification

Reject option (cont.)

We can achieve this by introducing a threshold θ and rejecting inputs $\mathbf{x}$ for which the largest of the posterior probabilities $p(\mathcal{C}_k|\mathbf{x})$ is less than θ



Setting $\theta = 1$ rejects all examples
Setting $\theta < 1/K$ rejects no examples

Inference and decision
Decision theory

Inference and decision

We have broken the classification problem down into two separate stages

- **Inference**: we use training data to learn a model for $p(\mathcal{C}_k|\mathbf{x})$
- **Decision**: we use $p(\mathcal{C}_k|\mathbf{x})$ to make optimal class assignments

Inference and decision - Generative models

1. Solve the inference problem of determining the class-conditional densities $p(\mathbf{x}|\mathcal{C}_k)$ for each class $\mathcal{C}_k$ individually
2. Separately infer the prior class probabilities $p(\mathcal{C}_k)$
3. Use Bayes' theorem to find posterior class probabilities $p(\mathcal{C}_k|\mathbf{x})$

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} \tag{6}$$

The denominator in Bayes' theorem can be found in terms of the quantities appearing in the numerator:

$$p(\mathbf{x}) = \sum_k p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k) \tag{7}$$

4. Use the posterior class probabilities with decision theory to determine class membership for each new input $\mathbf{x}$

We could also model the joint distribution $p(\mathbf{x}, \mathcal{C}_k)$ directly and then normalise to obtain posterior class probabilities for decision theory

Inference and decision - Discriminative models

1. Solve the inference problem of determining posterior class probabilities $p(\mathcal{C}_k|\mathbf{x})$
2. Use decision theory to assign each new $\mathbf{x}$ to one of the classes $\mathcal{C}_k$

Modelling the posterior class probabilities directly is **discriminative modelling**

Inference and decision - Discriminant function

There is an alternative, solve inference and decision together

- We learn a function that maps inputs into decisions
- Such a function is called a **discriminant function**

Typically, probabilities do not play any role in this approach

1. Find a discriminant function $f(\mathbf{x})$ which maps each input $\mathbf{x}$ directly onto a class label $\mathcal{C}_k$

Inference and decision (cont.)

Generative modelling is the most demanding approach

- It involves finding the joint distribution over both $\mathbf{x}$ and $\mathcal{C}_k$
- With high-dimensional inputs $\mathbf{x}$, we may need a large training set to be able to determine the class conditional densities accurately
- Class priors $p(\mathcal{C}_k)$ can often be estimated from fractions of the training data in each of the classes

There are also several advantages:

- A generative models allows to sample from it to generate synthetic data
- Because the marginal density of the data $p(\mathbf{x})$ can be calculated as

$$p(\mathbf{x}) = \sum_k p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)$$

we can detect new points that have low probability under the model[2]

[2] This is known as **outlier** or **novelty detection**

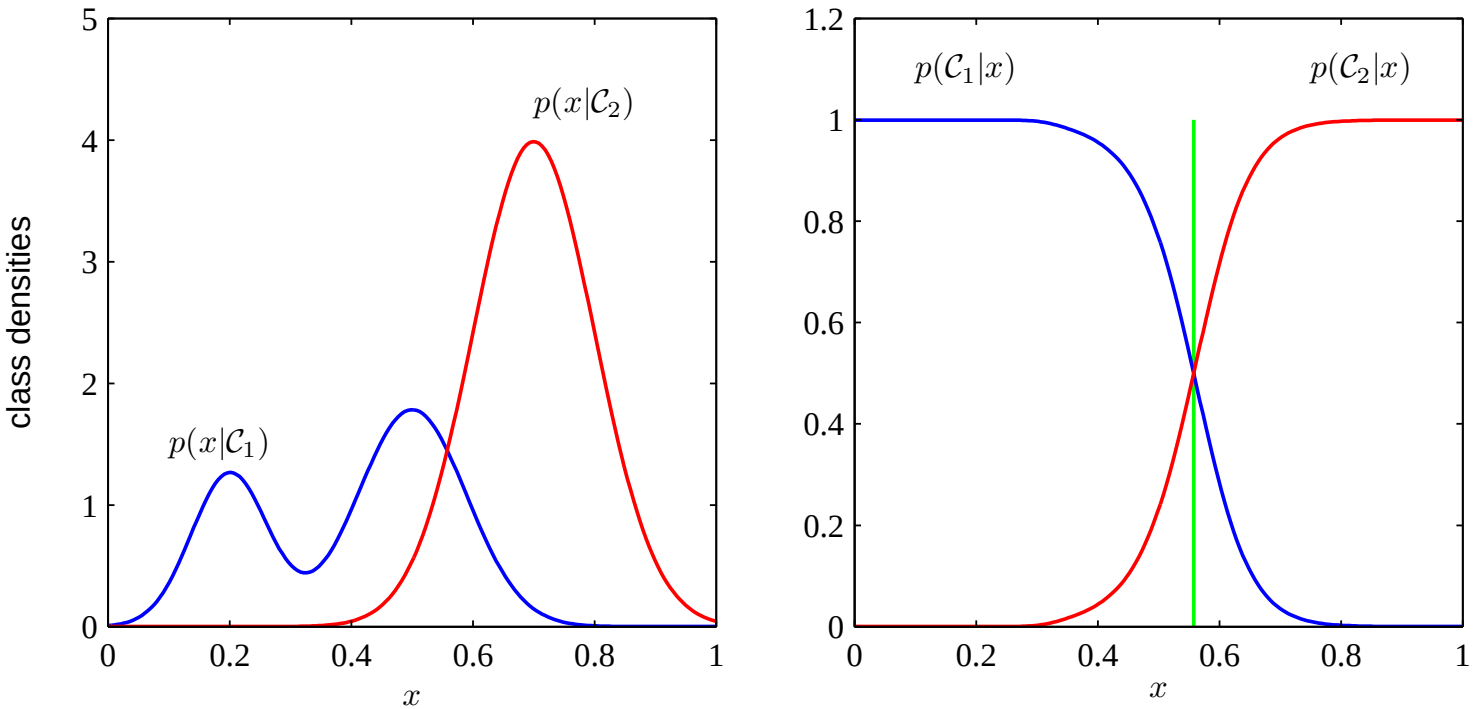
Inference and decision (cont.)

If we only wish to make classification decisions, then it is wasteful of resources, and excessively demanding of data, to find the joint distribution $p(\mathbf{x}, \mathcal{C}_k)$

- Especially because we only really need the posterior probabilities $p(\mathcal{C}_k|\mathbf{x})$, which can be obtained directly through a discriminative model
- Moreover, class conditional probabilities $p(\mathbf{x}|\mathcal{C}_k)$ may contain a lot of structure that has little effect on the posterior class probabilities $p(\mathcal{C}_k|\mathbf{x})$

Inference and decision (cont.)

The class-conditional densities $p(x|\mathcal{C}_k)$ for $K = 2$ classes having a single input variable x, together with the corresponding posterior probabilities $p(\mathcal{C}_k|x)$



The left-hand mode of the class-conditional probability $p(x|\mathcal{C}_1)$ has no effect on the corresponding posterior probabilities

Inference and decision (cont.)

An even simpler approach is the discriminative function in which we use training data to find a function $f(\mathbf{x})$ that maps each $\mathbf{x}$ directly onto a class label $\mathcal{C}_k$

- Inference and decision are combined into a single learning problem

With reference to the previous plot

- This corresponds to finding the value x shown by the vertical line
- This is the decision boundary giving minimum classification rate or equivalently minimum probability of misclassification

We have no access to posterior probabilities $p(\mathcal{C}_k|\mathbf{x})$ though

Loss functions for regression

Decision theory

Loss functions for regression

We discusses decision theory for classification problems, what for regression?

For regression problems (e.g., polynomial curve fitting), the decision stage consists of choosing a specific estimate $y(\mathbf{x})$ of the target $\mathbf{t}$ for each input $\mathbf{x}$

We can do this using a loss $L\Big(t, y(\mathbf{x})\Big)$, with the average or expected loss is

$$\mathbb{E}[L] = \int \int L\Big(t, y(\mathbf{x})\Big) p(\mathbf{x}, t) d\mathbf{x} dt \tag{8}$$

A common choice of loss function in regression problems is the squared loss

$$L(t, y(\mathbf{x})) = \Big(y(\mathbf{x}) - t\Big)^2$$

thus the expected loss can be written as

$$\mathbb{E}[L] = \int \int \Big(y(\mathbf{x}) - t\Big)^2 p(\mathbf{x}, t) d\mathbf{x} dt \tag{9}$$

Loss functions for regression (cont.)

$$\mathbb{E}[L] = \int \int \Big(y(\mathbf{x}) - t\Big)^2 p(\mathbf{x}, t) d\mathbf{x} dt$$

Our goal is to choose $y(\mathbf{x})$ so as to minimise $\mathbb{E}[L]$

Assume a completely flexible function $y(\mathbf{x})$, we can do this formally

$$\frac{\delta \mathbb{E}[L]}{\delta y(\mathbf{x})} = 2 \int \Big(y(\mathbf{x}) - t\Big) p(\mathbf{x}, t) dt = 0 \tag{10}$$

Solving for $y(\mathbf{x})$ and using sum and product rules of probability, we get

$$y(\mathbf{x}) = \frac{\int t p(\mathbf{x}, t) dt}{p(\mathbf{x})} = \int t p(t|\mathbf{x}) dt = \mathbb{E}_t[t|\mathbf{x}] \tag{11}$$

The quantity $\mathbb{E}_t[t|\mathbf{x}]$ is the conditional average of t conditioned on $\mathbf{x}$

Loss functions for regression (cont.)

$$\frac{\delta \mathbb{E}[L]}{\delta y(\mathbf{x})} = 2\int \Big(y(\mathbf{x}) - t\Big)p(\mathbf{x}, t)dt = 0 \quad \Longrightarrow \quad \cancel{2}\int \Big(y(\mathbf{x}) - t\Big)p(\mathbf{x}, t)dt = 0$$

$$\int y(\mathbf{x})p(\mathbf{x}, t)dt = \int tp(\mathbf{x}, t)dt \quad \Longrightarrow \quad y(\mathbf{x})\int p(\mathbf{x}, t)dt = \int tp(\mathbf{x}, t)dt$$

$$y(\mathbf{x}) = \frac{\int tp(\mathbf{x}, t)dt}{\int p(\mathbf{x}, t)dt} \quad \Longrightarrow \quad \frac{\int tp(\mathbf{x}, t)dt}{p(\mathbf{x})}$$

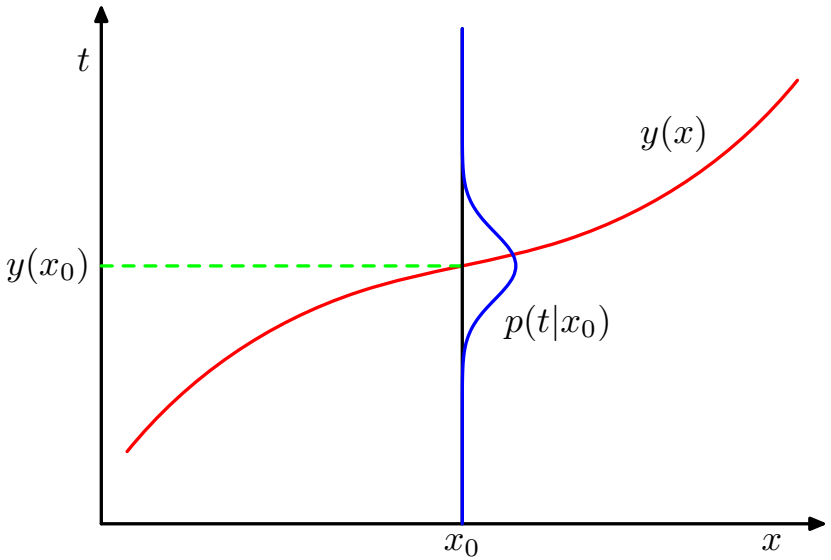
$$\frac{\int tp(t|\mathbf{x})p(\mathbf{x})dt}{p(\mathbf{x})} \quad \Longrightarrow \quad \frac{\cancel{p(\mathbf{x})}\int tp(t|\mathbf{x})dt}{\cancel{p(\mathbf{x})}}$$

$$\int p(t|\mathbf{x})tdt \quad = \quad \mathbb{E}_t[t|\mathbf{x}]$$

Loss functions for regression (cont.)

The **regression function** $y(\mathbf{x}) = \mathbb{E}_t[t|\mathbf{x}]$ minimises the expected squared loss

- The mean of the conditional distribution $p(t|\mathbf{x})$



The mean of the conditional distribution $p(t|\mathbf{x})$ and the regression function $y(\mathbf{x})$ are thus the same thing

Loss functions for regression (cont.)

We can also derive this result in a different way

- Start from the square term $\Big(y(\mathbf{x}) - t\Big)^2$

$$\begin{aligned}\Big(y(\mathbf{x} - t)^2\Big) &= \Big(y(\mathbf{x}) - \mathbb{E}_t[t|\mathbf{x}] + \mathbb{E}_t[t|\mathbf{x}] - t\Big)^2 \\ &= \Big(y(\mathbf{x}) - \mathbb{E}_t[t|\mathbf{x}]\Big)^2 \\ &+ 2\Big(y(\mathbf{x}) - \mathbb{E}_t[t|\mathbf{x}]\Big)\Big(\mathbb{E}_t[t|\mathbf{x}] - t\Big) \\ &+ \Big(\mathbb{E}_t[t|\mathbf{x}] - t\Big)^2\end{aligned}$$

Substituting into the loss function and integrating over t, we obtain

$$\mathbb{E}[L] = \int \Big(y(\mathbf{x}) - \mathbb{E}_t[t|\mathbf{x}]\Big)^2 p(\mathbf{x})d\mathbf{x} + \int \Big(\mathbb{E}_t[t|\mathbf{x}] - t\Big)^2 p(\mathbf{x})d\mathbf{x} \qquad (12)$$

which will be minimised when $y(\mathbf{x})$, in the 1-st term, equals $\mathbb{E}_t[t|\mathbf{x}]$

Loss functions for regression (cont.)

$$\mathbb{E}[L] = \int \Big(y(\mathbf{x}) - \mathbb{E}_t[t|\mathbf{x}]\Big)^2 p(\mathbf{x})d\mathbf{x} + \int \Big(\mathbb{E}_t[t|\mathbf{x}] - t\Big)^2 p(\mathbf{x})d\mathbf{x}$$

The second term is the variance of the distribution of t, averaged over $\mathbf{x}$

- It is the intrinsic variability of the target variable
- It can be seen as noise and cannot be reduced

Loss functions for regression (cont.)

As with classification, we can either determine appropriate probabilities and use them to make optimal decisions, or build models that make decisions directly

Again, three distinct approaches:

- Solve the inference problem of determining the joint density $p(\mathbf{x}, t)$, then normalise to find the conditional density $p(t|\mathbf{x})$ and finally marginalise to find the conditional mean $\mathbb{E}_t[t|\mathbf{x}]$
- Solve the inference problem of determining the conditional density $p(t|\mathbf{x})$ and then marginalise to find the conditional mean $\mathbb{E}_t[t|\mathbf{x}]$
- Find a regression function $y(\mathbf{x})$ directly from the training data

Loss functions for regression (cont.)

The squared loss is not the only possible choice of loss function for regression

- There are situations in which this loss can lead to poor results
- It is the case in which the conditional distribution $p(t|\mathbf{x})$ is multimodal, a situation that often arises in the solution of inverse problems

A generalisation of the loss function is the **Minkowski loss**, with expectation

$$\mathbb{E}[L_q] = \int \int \left|y(\mathbf{x}) - t\right|^q p(\mathbf{x}, t) d\mathbf{x} dt \tag{13}$$

For $q = 2$, the expected Minkowski loss reduces to the expected squared loss

- minimised by the conditional mean

For $q = 1$, the expected Minkowski loss is minimised by the conditional median
For $q \to 0$, the expected Minkowski loss is minimised by the conditional mode

Loss functions for regression (cont.)