Advanced Data Analytics
Advanced Evaluation

Ian T. Nabney

Overview

- Decision theory
- Cost and loss
- Evaluation measures

Decision theory

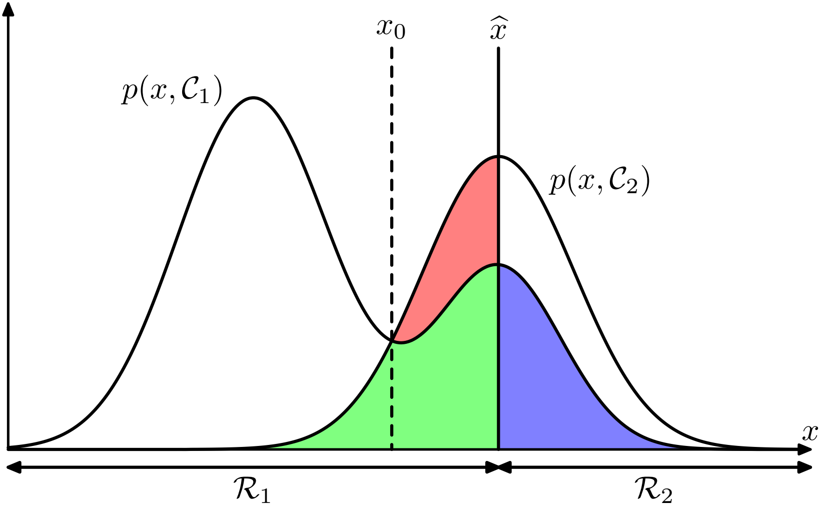
- Suppose we have an input vector $\mathbf{x}$ together with a corresponding vector $\mathbf{t}$ of target variables, and our goal is to predict $\mathbf{t}$ given a new value for $\mathbf{x}$. The joint probability distribution $p(\mathbf{x}, \mathbf{t})$ provides a complete summary of the uncertainty associated with these variables.
- In practical applications, we must often take a specific action based on our understanding of the values $\mathbf{t}$ is likely to take, and this aspect is the subject of decision theory.
- Consider, for example, 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 cancer or not. In this case, the input vector $\mathbf{x}$ is the set of pixel intensities in the image, and output variable t will represent the presence of cancer, which we denote by the class $\mathcal{C}_1$, or the absence of cancer, which we denote by the class $\mathcal{C}_2$.

University of BRISTOL

Minimising misclassification rate

- Suppose that our goal is simply to make as few misclassifications as possible. We need a rule that assigns each value of $\mathbf{x}$ to one of the available classes. Such a rule will divide the input space into regions $\mathcal{R}_k$ called decision regions, one for each class, such that all points in $\mathcal{R}_k$ are assigned to class $\mathcal{C}_k$.
- The boundaries between decision regions are called decision boundaries.
- A mistake occurs when an input vector belonging to class $\mathcal{C}_1$ is assigned to class $\mathcal{C}_2$ or vice versa. The probability of this occurring is given by

$$\begin{aligned} p(\mathrm{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\,\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\,\mathrm{d}\mathbf{x}. \qquad (1) \end{aligned}$$


x_0
$\widehat{x}$
$p(x, \mathcal{C}_1)$
$p(x, \mathcal{C}_2)$
x
$\mathcal{R}_1$
$\mathcal{R}_2$

Cost structure

- For many applications, our objective will be more complex than simply minimizing the number of misclassifications.
- In practice, false positive and false negative errors often incur different costs.
- Which cost is greater in each case?
  - Medical diagnostic tests: does X have leukaemia?
  - Loan decisions: approve mortgage for X?
  - Web mining: will X click on this link?
  - Promotional mailing: will X buy the product?

Loss functions

- We can formalize such issues through the introduction of a loss function, also called a cost function, which is a single, overall measure of loss incurred in taking any of the available decisions or actions. Our goal is then to minimize the total loss incurred.
- If model outputs estimates of posterior probabilities $P(\mathcal{C}_k|\mathbf{x})$, we can form a weighted sum to minimise expected cost.
- Let L_{kj} denote the cost of assigning an example to class $\mathcal{C}_j$ when it really belongs to class $\mathcal{C}_k$.
- Expected total cost of classifying to class $\mathcal{C}_j$ is $\sum_k L_{kj} P(\mathcal{C}_k|\mathbf{x})$. Choose j to minimise this.
- This avoids any question of rebalancing (and what is the right balance to use).

Example

Consider cost matrix

$$\begin{bmatrix} 0 & 500 \\ 1 & 0 \end{bmatrix}$$

- What sort of application might this be relevant to?
- Compute expected cost of classifying $\mathbf{x}$ as class $\mathcal{C}_1$:

University of BRISTOL

Example

Consider cost matrix

$$\begin{bmatrix} 0 & 500 \\ 1 & 0 \end{bmatrix}$$

- What sort of application might this be relevant to?
- Compute expected cost of classifying $\mathbf{x}$ as class $\mathcal{C}_1$:
- $$L_{11} P(\mathcal{C}_1|\mathbf{x}) + L_{21} P(\mathcal{C}_2|\mathbf{x}) \tag{2}$$
- Now you compute expected cost of classifying $\mathbf{x}$ as class $\mathcal{C}_2$:

Example

Consider cost matrix

$$\begin{bmatrix} 0 & 500 \\ 1 & 0 \end{bmatrix}$$

- What sort of application might this be relevant to?
- Compute expected cost of classifying $\mathbf{x}$ as class $\mathcal{C}_1$:
- $$L_{11} P(\mathcal{C}_1|\mathbf{x}) + L_{21} P(\mathcal{C}_2|\mathbf{x}) \tag{2}$$
- Now you compute expected cost of classifying $\mathbf{x}$ as class $\mathcal{C}_2$:
- $$L_{12} P(\mathcal{C}_1|\mathbf{x}) + L_{22} P(\mathcal{C}_2|\mathbf{x}) \tag{3}$$
- And the conclusion is:

Example

Consider cost matrix

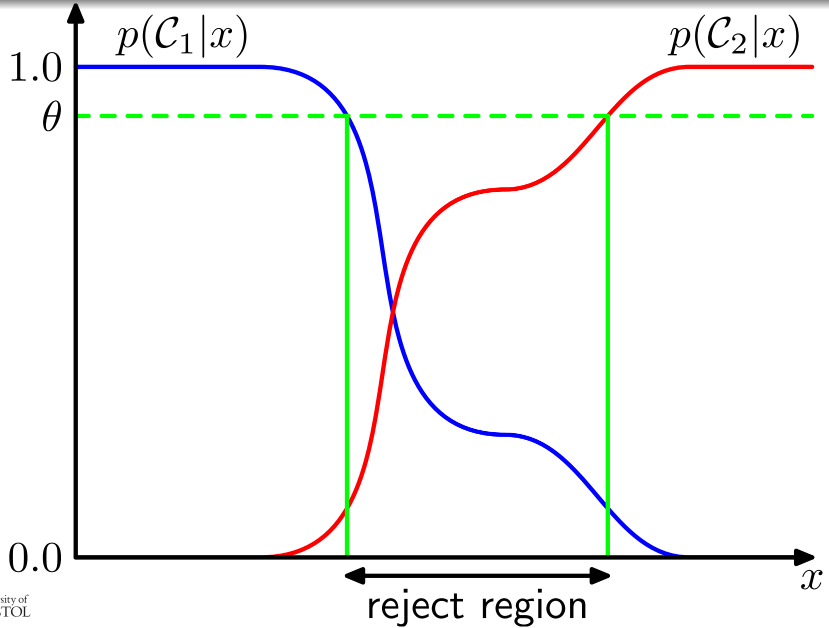
$$\begin{bmatrix} 0 & 500 \\ 1 & 0 \end{bmatrix}$$

- What sort of application might this be relevant to?
- Compute expected cost of classifying $\mathbf{x}$ as class $\mathcal{C}_1$:
- $$L_{11}P(\mathcal{C}_1|\mathbf{x}) + L_{21}P(\mathcal{C}_2|\mathbf{x}) \tag{2}$$
- Now you compute expected cost of classifying $\mathbf{x}$ as class $\mathcal{C}_2$:
- $$L_{12}P(\mathcal{C}_1|\mathbf{x}) + L_{22}P(\mathcal{C}_2|\mathbf{x}) \tag{3}$$
- And the conclusion is:
- Classify $\mathbf{x}$ as $\mathcal{C}_1$ unless $P(\mathcal{C}_2|\mathbf{x}) > 500P(\mathcal{C}_1|\mathbf{x})$.

Reject option

- Often classification errors arise from the regions of input space where the largest of the posterior probabilities $p(\mathcal{C}_k|\mathbf{x})$ is significantly less than unity, or equivalently where the joint distributions $p(\mathbf{x}, \mathcal{C}_k)$ have comparable values.
- These are the regions where we are relatively uncertain about class membership.
- In some applications, it is appropriate to avoid making decisions on the difficult cases in anticipation of a lower error rate on those examples for which a classification decision is made. This is known as the reject option.
- We can achieve this by introducing a threshold θ and rejecting those inputs $\mathbf{x}$ for which the largest of the posterior probabilities $p(\mathcal{C}_k|\mathbf{x})$ is less than or equal to θ.

Reject option schematic



Inference and decision

There are three distinct approaches to solving decision problems (in decreasing order of complexity)

1. First solve the inference problem of determining the class-conditional densities $p(\mathbf{x}|\mathcal{C}_k)$ for each class $\mathcal{C}_k$ individually. Also separately infer the prior class probabilities $p(\mathcal{C}_k)$. Then use Bayes' theorem in the form

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})} \tag{4}$$

   to find the posterior class probabilities $p(\mathcal{C}_k|\mathbf{x})$.

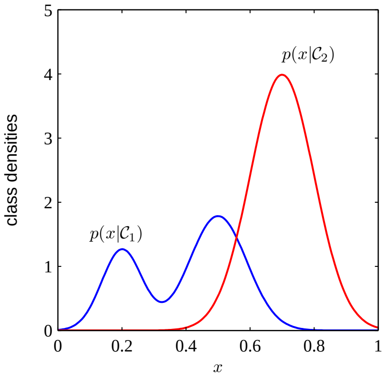
2. First solve the inference problem of determining the posterior class probabilities $p(\mathcal{C}_k|\mathbf{x})$, and then subsequently use decision theory to assign each new $\mathbf{x}$ to one of the classes. Approaches that model the posterior probabilities directly are called discriminative models.

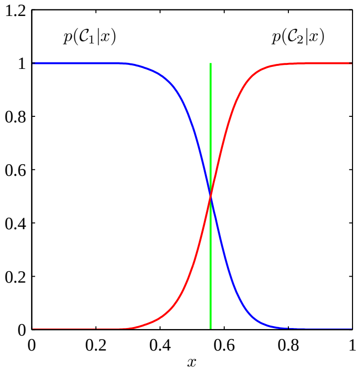
3. Find a function $f(\mathbf{x})$, called a discriminant function, which maps each input $\mathbf{x}$ directly onto a class label. In this case, probabilities play no role.

Comparing approaches

1. The most demanding because it involves finding the joint distribution over both $\mathbf{x}$ and $\mathcal{C}_k$. For many applications, $\mathbf{x}$ will have high dimensionality, and consequently we may need a large training set in order to be able to determine the class-conditional densities to reasonable accuracy.
2. If we only wish to make classification decisions, then it can be wasteful of computational resources, and excessively demanding of data, to find the joint distribution $p(\mathbf{x}, \mathcal{C}_k)$ when in fact we only really need the posterior probabilities $p(\mathcal{C}_k|\mathbf{x})$.
3. The goal is to find the decision boundary. We no longer have access to the posterior probabilities $p(\mathcal{C}_k|\mathbf{x})$.

Inference schematic





Evaluation measures

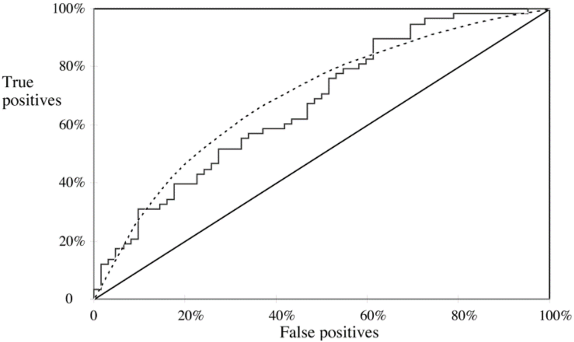
- We have seen how accuracy by itself may be a misleading or incomplete measure.
- If the classes are very imbalanced, then the default classifier has a very high accuracy.
- Accuracy takes no account of cost measures.
- We may want to impose a threshold on the output but have a range of possible thresholds to consider.
- For all these reasons, there are other evaluation measures used for classification tasks.

Confusion matrices

		Predicted class	
		Yes	No
Actual class	Yes	TP: True positive	FN: False negative
	No	FP: False positive	TN: True negative

- Machine learning algorithms usually minimise $FP + FN$.
- Direct marketing maximises TP.
- True positive rate $= TP/(TP + FN)$ also known as the sensitivity the probability of a positive test conditioned on being positive.
- False positive rate $= FP/(FP + TN)$.
- Specificity is the true negative rate $= TN/(TN + FP)$.

ROC curve

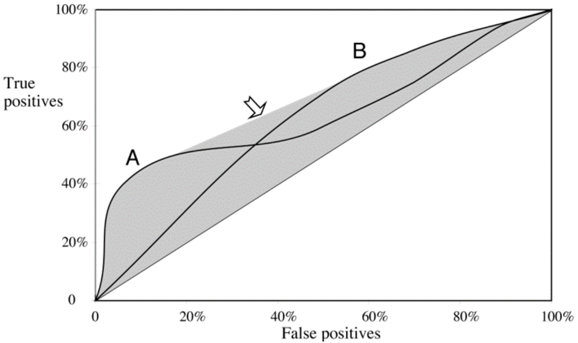


- Jagged curve: one set of test data
- Smooth curve: use cross-validation
- What does the straight line represent?

ROC curves for evaluation

- The area under the ROC curve (AUROC) can be interpreted as the probability that the classifier predicts the correct ordering of a pair of examples, one drawn from each class.
- The area can be used as a measure for comparing different classifiers over a range of decision thresholds (and costs).
- Simple method of getting an ROC curve using cross-validation:
  - Collect probabilities (scores) for instances in test folds
  - Sort instances according to probabilities
- `https://scikit-learn.org/stable/auto_examples/model_selection/plot_roc.html`

ROC curves for two schemes



- For a small, focused sample, use method A
- For a larger one, use method B
- In between, choose between A and B with appropriate probabilities

Convex hull models

- Given two learning schemes we can create a model that achieves any point on the convex hull!
- TP and FP rates for scheme A: t_1 and f_1
- TP and FP rates for scheme B: t_2 and f_2
- If scheme A is used to predict $100q$ % of the cases and scheme B for the rest, then
  - TP rate for combined scheme: $q \times t_1 + (1 - q) \times t_2$
  - FP rate for combined scheme: $q \times f_1 + (1 - q) \times f_2$

Evaluation measures for regression

- Assume target values $t_1, \ldots, t_N$ and predictions $y_1, \ldots, y_N$.
- The Mean squared error is

$$\frac{(t_1 - y_1)^2 + \cdots + (t_N - y_N)^2}{N} = \frac{\sum_{i=1}^{N}(t_i - y_i)^2}{N} \tag{5}$$

- This is easy to manipulate mathematically and has good properties relating to conditional mean.
- The root mean-squared error is measured in the same units as the target variable:

$$\sqrt{\frac{\sum_{i=1}^{N}(t_i - y_i)^2}{N}} \tag{6}$$

- The mean absolute error is less sensitive to outliers than the mean-squared error

$$\frac{|t_1 - y_1| + \cdots + |t_N - y_N|}{N} = \frac{\sum_{i=1}^{N}|t_i - y_i|}{N} \tag{7}$$

University of BRISTOL

Improvement on the mean

- These measures depend on the scaling of the target variable. Instead consider how much the scheme improves on simply predicting the average.
- The relative squared error is

$$\frac{\sum_{i=1}^{N}(t_i - y_i)^2}{\sum_{i=1}^{N}(\bar{t} - t_i)^2} \tag{8}$$

- The relative absolute error is

$$\frac{\sum_{i=1}^{N}|t_i - y_i|}{\sum_{i=1}^{N}|\bar{t} - t_i|} \tag{9}$$

- Want these values to be near zero. A value of 1 indicates a model no better than predicting the target mean (equivalent to default rule in classification).
- These measures give us an absolute evaluation that doesn't depend on variable scaling.

Overview

- Decision theory
- Cost and loss
- Evaluation measures