Ch.5 Classification and Clustering

Classification [Book, scattered]

With *discrete* response variables, do classification instead of regression.

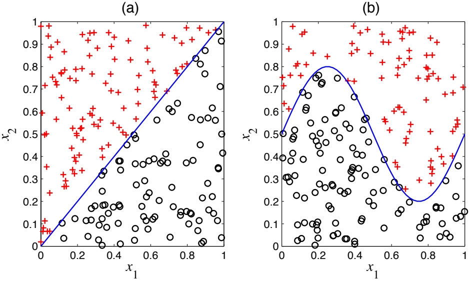

Figure : (a) Linear and (b) nonlinear decision boundaries separating 2 classes in the feature space (i.e. predictor space).

Classes are separated by decision boundaries , which can be linear or nonlinear. Linear classifiers only give linear decision boundaries, but nonlinear classifiers can give nonlinear decision boundaries.

In a parametric model, the model structure is specified by a number of model parameters (e.g. in linear regression, the regression coefficients are the parameters).

In a non-parametric model , the number of parameters is flexible and not fixed in advance, and may grow with the amount of training data.
We start with a simple non-parametric classifier: k-nearest neighbours.

5.1 k-nearest neighbour (kNN) classifier

Given training data $\{\mathbf{x}, y\}$ to build the model. The discrete response variable y can belong to class C_i, $i = 1, 2, \ldots, c$. Given a new feature vector $\mathbf{x}'$, predict the response y'.

Choose k, the number of nearest neighbours.
Find the k nearest neighbours in the training data $\{\mathbf{x}\}$, closest to $\mathbf{x}'$.
For these k nearest neighbours, look at their corresponding y values.

If class C_i occurs more frequently than the other classes in the k y-values, then let $y' = C_i$.

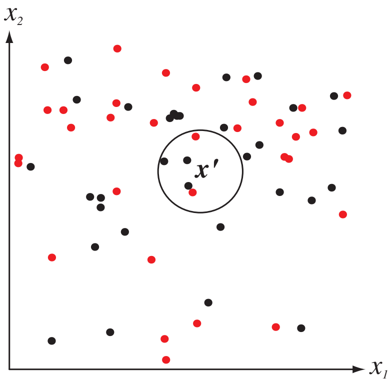

Figure : From test point $\mathbf{x}'$, grow spherical region until k nearest neighbours are enclosed. The class y' at $\mathbf{x}'$ is determined by the majority vote of the k nearest neighbours. (Duda et al., 2001, Fig.4.15)

One can even get a posterior probability for y'. If among the k y-values, n_1 of them belong to class C_1, n_2 to C_2, etc., then the probability of y' being in class C_1 is n_1/k, ..., in class C_i is n_i/k, etc. Classification then boils down to voting, i.e. choose the class C_i with the most "votes" from the k nearest neighbours.

Q1: A k nearest neighbour model is used to predict summer temperature (warm, normal or cool) from the spring climate conditions $\mathbf{x}$. With $k = 10$, for a test point $\mathbf{x}'$, the 10 nearest neighbours from the training data have 3 warm, 3 normal and 4 cool summers. What is the posterior probability of the coming summer being (a) warm, (b) normal and (c) cool?

Choice of k: If k is small, classification is sensitive to noise in the training data. Bigger $k =>$ less sensitivity to noise. But if k is too large $=>$ some of the neighbours are far from $\mathbf{x}'$.

How to choose the optimal k value?
Use only part (e.g. 80%) of the available data for model training, keep the remainder as validation data.
Build a series of models: e.g. model 1 with $k = 1$, model 2 with $k = 2$, ..., model K with $k = K$.
Test the model performance using the independent validation data.
The model with the fewest classification errors gives the optimal choice for k.

Pros and cons of the kNN classifier:
Pros: Nonlinear classifier. Simple in concept.

Cons: (a) Need all training data to specify the model $=>$ needs lots of memory and computationally slow for datasets with many samples. (b) When the dimension of $\mathbf{x}$ is not small, the k nearest neighbours can be far away.
Overall, not a good method.

5.2 Conditional probabilities and Bayes’ theorem [Book, Sec.1.5; Bishop (2006, Sect.1.2)]

Let X and Y be discrete variables. E.g. $X =$ ‘warm’, ‘normal’ or ‘cold’; $Y =$ ‘drought’ or ‘no drought’. $P(X, Y)$ is the joint probability. E.g. $P(\text{cold, drought}) =$ probability of having both cold and drought conditions.

Two rules of probability:

$$P(X) = \sum_{Y} P(X, Y), \quad \text{(sum rule)} \tag{1}$$

$$P(X, Y) = P(Y|X)P(X) = P(X|Y)P(Y), \quad \text{(product rule)} \tag{2}$$

where $P(X)$ is the marginal probability, and $P(Y|X)$ is the conditional probability, i.e. the probability of Y given X. E.g. $P(\text{drought}|\text{warm})$ = probability of drought given it is warm.

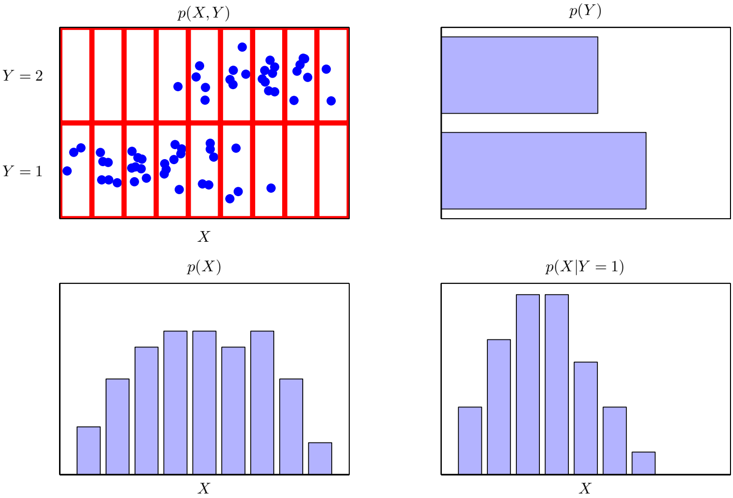

Figure 1.11 An illustration of a distribution over two variables, X, which takes 9 possible values, and Y, which takes two possible values. The top left figure shows a sample of 60 points drawn from a joint probability distribution over these variables. The remaining figures show histogram estimates of the marginal distributions $p(X)$ and $p(Y)$, as well as the conditional distribution $p(X|Y=1)$ corresponding to the bottom row in the top left figure. [From Bishop, 2006]

E.g. A meteorologist categories X, the air pressure in the morning, as L (low) or NL (non-L). He also categories Y, the occurrence of tornadoes in the afternoon as T (tornadoes) or NT (non-T).
From 100 days of observations, he found 20 days of L and T, 5 days of NL and T, 10 days of L and NT, and 65 days of NL and NT.
Find $P(X, Y)$, $P(X)$ and $P(Y)$.

	$P(X,Y)$		
Y:	X: L	NL	$P(Y)$
T	20/100	5/100	25/100
NT	10/100	65/100	75/100
$P(X)$	30/100	70/100	

Q2: Compute the 2×2 table of $P(Y|X)$ for the tornado problem.

—

From (2)/$P(X)$, we get Bayes' theorem:

$$\boxed{P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}}. \tag{3}$$

Substituting (2) into (1) gives

$$P(X) = \sum_Y P(X|Y)P(Y), \tag{4}$$

so Bayes' theorem can be written as

$$\boxed{P(Y|X) = \frac{P(X|Y)P(Y)}{\sum_Y P(X|Y)P(Y)}}. \tag{5}$$

$\sum_Y$Eq.(5) gives

$$\sum_Y P(Y|X) = 1. \tag{6}$$

Bayes' theorem, (Reverend Thomas Bayes, 1702–1761), plays a central role in modern statistics (Jaynes, 2003).
Bayesians describe probabilities in terms of beliefs and degrees of uncertainty, similar to how the general public uses probability.

E.g., a fan prior to the start of a sports tournament asserts team A has a probability of 60% for winning the tournament. After a loss, the fan modifies the winning probability to 30%.
Bayes' theorem provides formula for modifying prior probability $P(Y)$ in view of new data X.

Q3: Suppose a test for a toxin in a lake gives the following results: If a lake has the toxin, the test returns a positive result 99% of the time. If a lake does not have the toxin, the test still returns a positive result 2% of the time. Suppose only 5% of the lakes contain the toxin. What is the probability that a positive test result for a lake turns out to be a false positive?

—

More generally, e.g. a meteorologist wants to classify the approaching weather state as class C_i, $(i = 1, \ldots, k)$, e.g. sunny, cloudy, rainy, snowy, etc.
Assume some *a priori probability* (or simply *prior probability*) $P(C_i)$.

He gets meteorological observations $\mathbf{x}$ (continuous variables) at 6 a.m. The meteorologist would like to obtain an *a posteriori probability* (or simply *posterior probability*) $P(C_i|\mathbf{x})$, i.e. the conditional probability of having weather class C_i on that day given the 6 a.m. $\mathbf{x}$ data.

The joint probability density $p(C_i, \mathbf{x})$ is the probability density that an event belongs to class C_i and has value $\mathbf{x}$.
Notation: p denotes a probability density; P for probability.

The joint probability density can be written as

$$p(C_i, \mathbf{x}) = P(C_i|\mathbf{x})p(\mathbf{x}), \tag{7}$$

with $p(\mathbf{x})$ the probability density of $\mathbf{x}$. Alternatively,

$$p(C_i, \mathbf{x}) = p(\mathbf{x}|C_i)P(C_i), \tag{8}$$

with $p(\mathbf{x}|C_i)$, the conditional probability density of $\mathbf{x}$, given that the event belongs to class C_i. Equating the right hand sides of these 2 eqns. gives *Bayes' theorem*:

$$\boxed{P(C_i|\mathbf{x}) = \frac{p(\mathbf{x}|C_i)P(C_i)}{p(\mathbf{x})}}. \tag{9}$$

Eq.(4) generalizes to

$$p(\mathbf{x}) = \sum_i p(\mathbf{x}|C_i)P(C_i). \tag{10}$$

Eq.(9) becomes

$$\boxed{P(C_i|\mathbf{x}) = \frac{p(\mathbf{x}|C_i)P(C_i)}{\sum_i p(\mathbf{x}|C_i)P(C_i)}}, \tag{11}$$

where the denominator is a normalization factor for the posterior probabilities to sum to unity.

Bayes' theorem says that the posterior probability $P(C_i|\mathbf{x})$ is simply $p(\mathbf{x}|C_i)$ (the *likelihood* of $\mathbf{x}$ given the event is of class C_i) multiplied by the prior probability $P(C_i)$, and divided by a normalization factor.

If we have a *continuous* variable w (instead of discrete variable C_i), then Bayes' theorem (9) becomes:

$$\boxed{p(w|\mathbf{x}) = \frac{p(\mathbf{x}|w)p(w)}{p(\mathbf{x})}}. \tag{12}$$

5.3 Logistic regression [Bishop (2006, pp.197-198, 204-206, 209, 186-187), Book, pp.89-90, 174-176]

The classical approach to classification is by discriminant analysis, where the feature space (i.e. $\mathbf{x}$ space) is divided by decision boundaries into decision regions $R_1, \ldots, R_k$ — if a feature vector lands within R_i, the classifier will assign the class C_i. R_i may be composed of several disjoint regions, all of which are assigned the class C_i.

In *linear* discriminant analysis (LDA), decision boundaries are linear. LDA can be solved by various methods, e.g. least squares, Fisher's linear discriminant, etc. (Bishop, 2006, Sect.4.1).

LDA has been surpassed by newer methods like logistic regression (a linear classifier despite "regression" in the name).

Logistic regression arises naturally from Bayes' theorem. Start with two classes C_1 and C_2. Recall Bayes' theorem (11):

$$\begin{aligned}
P(C_1|\mathbf{x}) &= \frac{p(\mathbf{x}|C_1)P(C_1)}{p(\mathbf{x}|C_1)P(C_1)+p(\mathbf{x}|C_2)P(C_2)} && (13)\\
&= \frac{1}{1+\frac{p(\mathbf{x}|C_2)P(C_2)}{p(\mathbf{x}|C_1)P(C_1)}} && (14)\\
&= \frac{1}{1+\mathrm{e}^{-u}}, && (15)
\end{aligned}$$

with

$$-u = \ln\left[\frac{p(\mathbf{x}|C_2)P(C_2)}{p(\mathbf{x}|C_1)P(C_1)}\right]. \qquad (16)$$

The logistic sigmoidal function $\sigma(u)$

$$\sigma(u) = \frac{1}{1+\mathrm{e}^{-u}}. \qquad (17)$$

If we further assume that the two classes $p(\mathbf{x}|C_i)$ both have Gaussian distributions, one can show that

$$P(C_1|\mathbf{x}) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x} + b). \tag{18}$$

The constant parameter b can be dropped if we add 1 as the first element of $\mathbf{x}$ and add $w_0(= b)$ as the first element of $\mathbf{w}$, hence

$$P(C_1|\mathbf{x}) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}). \tag{19}$$

$$P(C_2|\mathbf{x}) = 1 - P(C_1|\mathbf{x}) = 1 - \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x}). \tag{20}$$

These 2 eqns. give the logistic regression model, but need to solve for the weights $\mathbf{w}$. Use maximum likelihood to find the optimal weights.

E.g. the dataset has 3 observations for $\mathbf{x}$, namely $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ and for the class C_i, which happens to be $\{C_2, C_2, C_1\}$. The likelihood function is

$$P(\{C_2, C_2, C_1\}|\mathbf{w}) = P(C_2|\mathbf{x}_1)P(C_2|\mathbf{x}_2)P(C_1|\mathbf{x}_3). \tag{21}$$

Find $\mathbf{w}$ which maximizes this likelihood function.

Multiclass (or multinomial) logistic regression is used when more than 2 classes, i.e. C_i $(i = 1, \dots, k)$, with

$$P(C_i|\mathbf{x}) = \frac{\exp(\mathbf{w}_i^{\mathrm{T}}\mathbf{x})}{\sum_{j=1}^{k} \exp(\mathbf{w}_j^{\mathrm{T}}\mathbf{x})}. \tag{22}$$

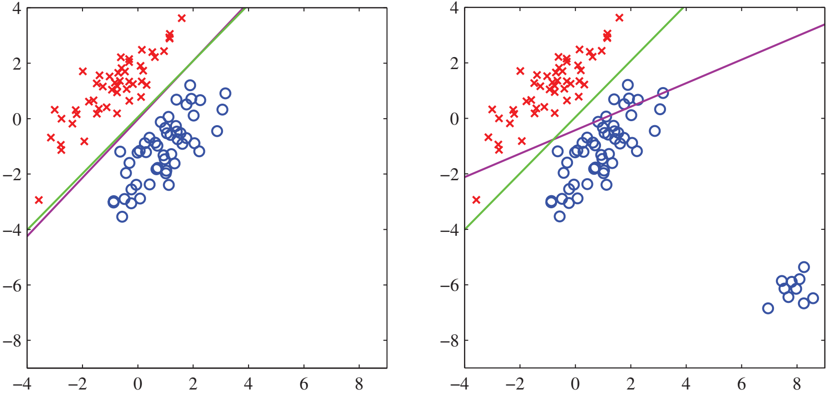

Figure 4.4 The left plot shows data from two classes, denoted by red crosses and blue circles, together with the decision boundary found by least squares (magenta curve) and also by the logistic regression model (green curve), which is discussed later in Section 4.3.2. The right-hand plot shows the corresponding results obtained when extra data points are added at the bottom left of the diagram, showing that least squares is highly sensitive to outliers, unlike logistic regression. [from Bishop (2006)]

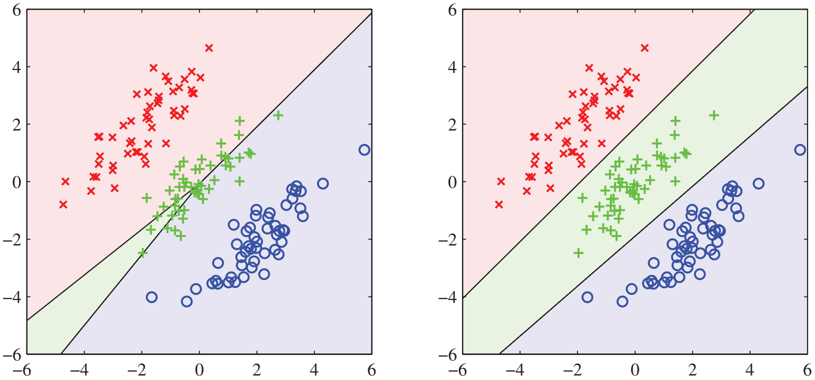

Figure 4.5 Example of a synthetic data set comprising three classes, with training data points denoted in red ($\times$), green ($+$), and blue ($\circ$). Lines denote the decision boundaries, and the background colours denote the respective classes of the decision regions. On the left is the result of using a least-squares discriminant. We see that the region of input space assigned to the green class is too small and so most of the points from this class are misclassified. On the right is the result of using logistic regressions as described in Section 4.3.2 showing correct classification of the training data. [From Bishop (2006)]

Matlab codes for classification:

k-nearest neighbour classifier:
`http://www.mathworks.com/help/stats/classificationknnclass.html`

(Multinomial) logistic regression:
`http://www.mathworks.com/help/stats/mnrfit.html`

References:

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer, New York.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification.* Wiley, New York, 2nd edition.

Jaynes, E. T. (2003). *Probability theory: the logic of science.* Cambridge Univ. Pr., Cambridge. Bretthorst, G.L. (editor).