Introduction to Statistical Pattern Recognition and Optimisation

Hiroshi Shimodaira*

January-March 2020

When we are trying to make decisions, uncertainty arises due to:

- Inaccurate or incomplete information about the situation (for example due to noisy or inaccurate sensors).
- A lack of complete knowledge about the situation (for example we may only have information arising from the available sensors, perhaps missing out on other sources of information).

We must assume that only uncertain and partial information is available, and the system must make decisions taking this into account. The mathematics of probability provides the way to deal with uncertainty, and tells us how to update our knowledge and beliefs if new information becomes available. This chapter introduces the use of probability and statistics for pattern recognition and learning.

5.1 A simple example

Consider the problem of determining the sex of fish. We could carefully examine the fish, which would be slow and tedious. For many species of fish it is known that male fish tend to be longer than female fish. It is much easier and faster to measure the length of a fish than to determine its sex directly. Thus we would like a system to determine the sex of fish from their length, as accurately as possible.[1]

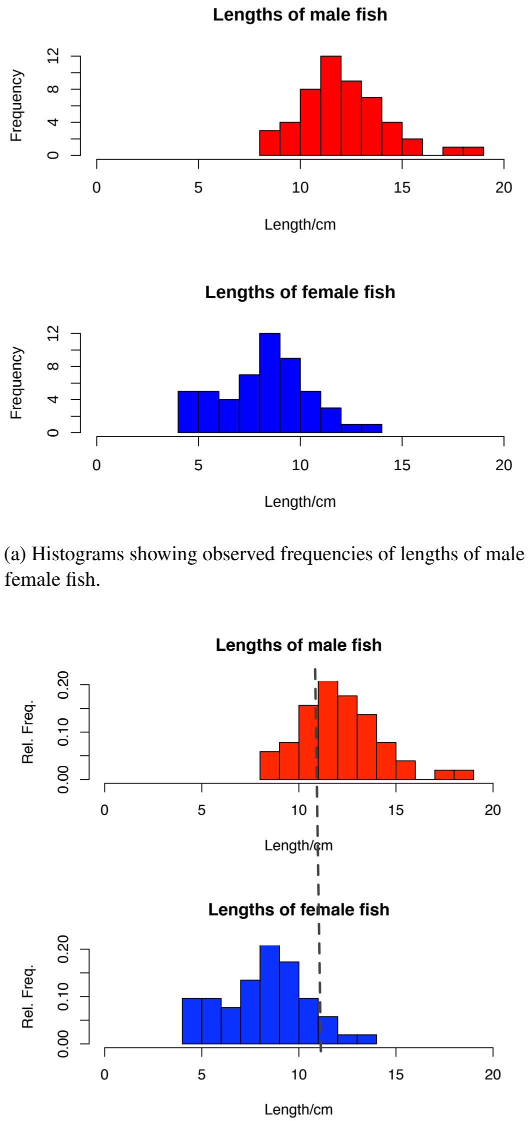
We will consider the following scenario: we have a set of *labelled data* that we can use as a training set. This is a set of measurements of the length of each fish, together with the class label (male or female). In Figure 5.1a we plot this dataset as two histograms, one for for each class (sex), showing the number of fish of each length in each class. The length data comes as integer values (cm); later we'll look at directly modelling continuous valued data.

We now have four new, unlabelled examples with lengths 5 cm, 9 cm, 12 cm, 16 cm. How do we classify each of these?

- 5 cm: It seems clear that it should be classified female, since we have never seen a male fish this short. (But can we be *sure* that we will *never* see a male fish of this length?)

This note is heavily based on notes inherited from Steve Renals and Iain Murray.

[1] This is based on a real problem; there is a classic paper (R. Redner and H. Walker, "Mixture densities, maximum likelihood, and the EM algorithm," SIAM Rev. 26 (1984), pp.195–239.) which uses as an example the determination of the sex of halibut using their length as the feature, based on data from the Seattle Fisheries Commission.



(a) Histograms showing observed frequencies of lengths of male and female fish.

(b) Histograms showing relative frequencies of lengths of male and female fish. The dotted line shows a possible decision boundary between male and female, based on the length.

Figure 5.1: Histograms showing lengths of male and female fish.

- 16 cm: Likewise, it seems safe to classify the 16 cm fish as male

- 9 cm: We have observed both male and female fish that are 9 cm long, but about twice as many females of that length as males. So it would seem sensible to classify that fish as female.

- 12 cm: We have also observed both male and female fish that are 12 cm long, but more male than female, so it seems to make sense to classify that fish as male.

For the 9 cm and 12 cm cases (in particular), it is not possible to make an unambiguous classification. The length of the fish gives us evidence whether it is male or female, but the length alone is not enough information to be sure.

Rather than plotting the histograms with the observed frequencies for each length, we could plot them using the *relative frequencies*—the proportions of male and female fish that are each length (Figure 5.1b).

We can assign a particular value of the length to be a *decision boundary*—that value of the length at which we assign all shorter fish to be female and longer ones to be male. A possible decision boundary is marked by the dotted line in Figure 5.1b. Intuitively it seems like that this is a good decision point, based on minimising the number of misclassifications in the training set. While we use the training set to determine the decision boundary, we are most interested in classifying new data (a test set).

We can think of the relative frequencies as *probability estimates*; estimates of the probability of the length given that the fish is male (or female). If we call the length X and the class S, then we can write that the relative frequencies are estimates of $P(X=\ell \mid S=\text{male})$ and $P(X=\ell \mid S=\text{female})$: the probability that the length X has value ℓ given that the fish is male (female).

Question: Imagine that you know that there are 10 times as many male fish as female fish. Would you still make the same classifications?

5.2 Probability refresher

Before jumping into the use of probabilities for pattern recognition and learning, let's briefly revise basic probability.

Some events are *deterministic*: if you drop an apple, it will hit the floor (event A *causes* event B).

Some events are random. They may be intrinsically random (such as radioactive decay) or they may be deterministic, but we do not have enough information (or it is prohibitive to do the computations). For example tossing a coin is a deterministic physical process: if we had enough information (force of the toss, weight of the coin, etc.) then it would be possible to compute the outcome of the toss precisely (using Newtonian mechanics). Forecasting the weather is another example in this category, but much more complex!

Probabilistic models are the correct model to use when acting under uncertainty: we never have the *whole truth*. We can use probability theory to make decisions based on what we know.

5.2.1 What is probability?

This turns out to be a deep and controversial question! Indeed, there is still debate about the correct interpretation of probability. There are two basic positions:

- Probability is a *frequency limit*: e.g., tossing a fair coin N times, $n(\text{heads})/N \to 1/2$ as $N \to \infty$;

- Probability is a *degree of belief*: e.g., given a set of symptoms a doctor may believe you a have a particular disease with probability $P(\text{disease} \mid \text{symptoms})$;

It turns out that it is possible to derive the mathematics of probability starting from a small set of axioms.

Sample space

A *trial* is the basic event which we want to model (e.g., toss of a coin, roll of a dice). A *Sample space* — set of all possible outcomes of a trial (e.g., $\{1, 2, 3, 4, 5, 6\}$ in the case of rolling a dice).

If the outcome of an experiment is A, then the complement of A is written $\overline{A}$ – 'not A', the set of all other outcomes. The sample space is $\{A, \overline{A}\}$. In this case we consider A as representing an event that may be true or false, and we have a sample space {true, false}, or $\{1, 0\}$.

Symmetry

We would like to assign probabilities to events in sample space, such that equivalent (symmetric, interchangeable) outcomes should be assigned the same probability (e.g., tossing a fair coin). And we would like to constrain the total probability of all outcomes to be 1. If a sample space is composed of N symmetric events (e.g., rolling a fair dice), then the probability of each event in the sample space should be $1/N$.

Independence

Two events A and B are *independent* if the outcome of A does not influence the outcome of B (and vice-versa). The sample space of two independent events is their Cartesian Product.

5.2.2 Boxes example (Bishop, 2006)

We have two boxes, 1 red and 1 green. In the red box we have 2 apples and 6 oranges. In the green box we have 3 apples and 1 orange. Suppose we randomly choose a box, and from the chosen box we randomly choose a piece of fruit. We choose the red box 40% of the time, and the green box 60% of the time.

In this example we have two *random variables*:

- B the identity of the box, which can take two values r (red) or g (green)

- F the type of fruit, with two possible values a (apple) and o (orange)

We can define the probability of an event as the fraction of times that event occurs (as $N \to \infty$). We can write the probability of selecting the red box as $P(B=r) = 4/10$, and the probability of selecting the green box as $P(B=g) = 6/10$. We only have two boxes and we must choose one of them (B can only take values r or g), so

$$P(B=r) + P(B=g) = 1.$$

We'll come back to this example.

5.2.3 Joint and Conditional Probability

Assume we have two random variables X and Y. X can take one of I values $x_1, \ldots, x_i, \ldots, x_I$ and Y can take one of J values $y_1, \ldots, y_j, \ldots, y_J$. Consider N trials where we sample the values of X and Y, and let n_{ij} be the number of times $X = x_i$ and $Y = y_j$. If we consider the limit $N \to \infty$, then we can define the *joint probability* that X has value x_i and Y has value y_j as:

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N}\,.$$

Sometimes, to avoid clutter, we write $P(x_i, y_j)$. The ordering of the terms is irrelevant here: $P(X = x_i, Y = y_j) = P(Y = y_j, X = x_i)$.

Let n_i be the number of times $X = x_i$, irrespective of the value of Y. Then we can write:

$$P(X = x_i) = \frac{n_i}{N}\,,$$

which can also be abbreviated as $P(x_i)$. From the definition of n_i,

$$n_i = \sum_{j=1}^{J} n_{ij}\,.$$

Therefore we can write

$$P(X = x_i) = \sum_{j=1}^{J} P(X = x_i, Y = y_j)\,, \tag{5.1}$$

which is called the *law of total probability* or the *sum rule*.

The *conditional probability* $P(X = x_i \mid Y = y_j)$, read as the probability that $X = x_i$ given that $Y = y_j$, is obtained by only considering events when $Y = y_j$, and is the proportion of the time when $X = x_i$ in that case:

$$P(X = x_i \mid Y = y_j) = \frac{n_{ij}}{n_j}\,.$$

Now, since

$$\frac{n_{ij}}{N} = \frac{n_{ij}}{n_j}\frac{n_j}{N}\,,$$

we may write the *product rule* which relates the joint probability to the conditional probability:

$$P(X = x_i, Y = y_j) = P(X = x_i \mid Y = y_j)\, P(Y = y_j)\,. \tag{5.2}$$

If we want to write the distribution for an arbitrary value of the random variable X we can write $P(X)$. Using this more compact notation we can write the sum and product rules as:

$$P(X) = \sum_{Y} P(X, Y) \tag{5.3}$$

$$P(X, Y) = P(X|Y)\, P(Y) = P(Y|X)\, P(X)\,. \tag{5.4}$$

You can read $P(X, Y)$ as 'the probability of X *and* Y', read $P(Y|X)$ as 'the probability of Y *given* X', and read $P(X)$ as 'the probability of X'. $\sum_Y$ refers to the sum over all values that Y can take.

5.2.4 Bayes' Theorem

Re-expressing Equation (5.4), we can write:

$$P(Y|X) = \frac{P(X|Y)\, P(Y)}{P(X)}\,. \tag{5.5}$$

This is known as *Bayes' Theorem*. It is extremely useful for computing $P(Y|X)$ when $P(X|Y)$ is known. Bayes' theorem and the law of total probability are at the centre of statistical pattern recognition and machine learning.

Using the law of total probability (sum rule) we can expand the denominator as:

$$P(X) = \sum_{Y} P(X|Y)\, P(Y)\,,$$

and we can write Bayes' theorem as

$$P(Y|X) = \frac{P(X|Y)\, P(Y)}{\sum_Y P(X|Y)\, P(Y)}\,. \tag{5.6}$$

5.2.5 Boxes example continued

In this example we had:

$$\begin{aligned} P(B = r) &= 4/10 \\ P(B = g) &= 6/10\,. \end{aligned}$$

The probability of picking an apple from the red box (conditional probability of picking an apple given that red box was chosen) is the fraction of apples in the red box (1/4) and is written as $P(F = a \mid B = r)$. We can write the set of conditional probabilities of picking a type of fruit given a box:

$$\begin{aligned} P(F = a \mid B = r) &= 1/4 \\ P(F = o \mid B = r) &= 3/4 \\ P(F = a \mid B = g) &= 3/4 \\ P(F = o \mid B = g) &= 1/4\,. \end{aligned}$$

As a check, we can verify that each conditional distribution is normalised:
$P(F = a \mid B = r) + P(F = o \mid B = r) = 1$ and $P(F = a \mid B = g) + P(F = o \mid B = g) = 1$.

We can use the law of total probability to evaluate the overall probability of choosing an apple:

$$\begin{aligned} P(F = a) &= P(F = a \mid B = r)\, P(B = r) + P(F = a \mid B = g)\, P(B = g) \\ &= 1/4 \cdot 4/10 + 3/4 \cdot 6/10 \\ &= 22/40 = 11/20\,. \end{aligned}$$

And the probability of choosing an orange is:

$$P(F = o) = 1 - P(F = a) = 9/20\,.$$

Now suppose we are told that an apple was chosen, but we don't know which box it came from. We can use Bayes' theorem to evaluate the conditional probability that the red box was chosen, given that an apple was picked:

$$\begin{aligned} P(B=r \mid F=a) &= \frac{P(F=a \mid B=r)\, P(B=r)}{P(F=a)} \\ &= \frac{1/4 \cdot 4/10}{11/20} \\ &= 2/11\,. \end{aligned}$$

So, the probability that red box was chosen given that an apple was picked is 2/11, and

$$P(B=g \mid F=a) = 1 - P(B=r \mid F=a) = 1 - 2/11 = 9/11\,,$$

the probability that green box was chosen given that an apple was picked is 9/11.

We can interpret Bayes' theorem as follows. Without having picked any fruit, the best information we have about the probabilities of the two boxes are given by the *prior probabilities* $P(B)$. Once we have picked a fruit, then we have some additional information and we can use Bayes' theorem to compute the *posterior probabilities* $P(B|F)$. Without having observed any fruit, the probability of choosing the green box is 6/10. If an apple is observed we can incorporate this information, finding that the posterior probability is 9/11. Observing the apple makes it more probable that the box we selected was the green one.

5.3 Bayes' Theorem and Pattern Classification

Bayes' theorem is at the heart of statistical pattern recognition. Let's look at how, in general terms, we can express a pattern classification problem using Bayes' theorem.

Consider a pattern classification problem in which there are K classes. Let C denote the class, taking values $1, \ldots, K$. The observed input data, which is a D-dimensional feature vector, is denoted by X. Once the training set is used to train the classifier, a new, unlabelled data point $\mathbf{x}$ is observed. To make a classification we could compute the *posterior probabilities* (also called *a posteriori* probabilities) $P(C=k|X=\mathbf{x})$, for every class $k = 1, \ldots, K$; we can then classify $\mathbf{x}$ by assigning it to the class with the highest posterior probability, k_{max}. We can write this operation as:

$$k_{max} = \arg\max_{k\in\{1,\ldots,K\}} P(C=k|X=\mathbf{x})\,. \tag{5.7}$$

This procedure is sometimes called MAP (maximum a posteriori) decision rule. The max operator returns a probability that is the maximum value of $P(C=k|X=\mathbf{x})$ over all values of C; the arg max operator returns the argument (the value of k_{max}) corresponding to the maximum probability. More compactly, we can write $P(C=k|X=\mathbf{x})$ as $P(C_k|\mathbf{x})$.[2]

MAP classification of $\mathbf{x}$ requires estimates of the conditional probability of each class, k, given $\mathbf{x}$. We can re-express the required conditional probabilities using Bayes' theorem:

$$P(C_k|\mathbf{x}) = \frac{P(\mathbf{x}|C_k)\, P(C_k)}{P(\mathbf{x})}\,. \tag{5.8}$$

We have expressed the posterior probability $P(C_k|\mathbf{x})$ as a product of:

[2] We will simply use C_k to denote $C = k$ for the ease of readability.

- the *likelihood*[3] of the class k, $P(\mathbf{x}|C_k)$, given data $\mathbf{x}$;
- the *prior probability*, $P(C_k)$, of the class k.

There is also the denominator, $P(\mathbf{x})$, to consider. We can compute this term using the law of total probability:

$$P(\mathbf{x}) = \sum_{\ell} P(\mathbf{x}|C_\ell)\, P(C_\ell)\,.$$

However, because $P(\mathbf{x})$ is the same for all classes, we do not need to consider it when finding the most probable class, so that $P(C_k|\mathbf{x})$ is proportional to $P(\mathbf{x}|C_k)\, P(C_k)$, i.e.

$$P(C_k|\mathbf{x}) \propto P(\mathbf{x}|C_k)\, P(C_k)\,. \tag{5.9}$$

If we return to the fish example, we can see why re-expressing the posterior probability using Bayes' theorem is a useful thing to do.

5.4 Fish example redux

We have 200 examples of fish lengths, 100 male and 100 female. Let class $C = M$ represent male, and $C=F$ represent female. Remember that we are assuming fish come in integer lengths—or, more realistically, if a fish has a length between 9.5 and 10.5, then $X = 10$. Let the number of male and female fish from with length x be given by $n_M(x)$ and $n_F(x)$, and the total number of fish in each class k are denoted N_M and N_F. These counts are given in table 5.1 for our example dataset.

We can estimate the likelihoods $P(x|M)$ and $P(x|F)$ as the counts in each class for length x divided by the total number of examples in that class:

$$\begin{aligned} P(x|M) &\approx \frac{n_M(x)}{N_M} \\ P(x|F) &\approx \frac{n_F(x)}{N_F}\,. \end{aligned}$$

To estimate these likelihoods, we consider the fish in each class separately (probability estimates are conditional on the class). We obtain estimates of $P(x|M)$ and $P(x|F)$ (and NOT estimates of $P(M|x)$ and $P(F\,|\,x)$). Thus we have estimated the likelihoods of the length given each class using relative frequencies (using the training set of 100 examples from each class). These are also tabulated in table 5.1. (These are probability *estimates* since N_M and N_F are finite, which is always the case for the real world.)

We can use Bayes' theorem to estimate the posterior probabilities of each class given the data:

$$\begin{aligned} P(M|x) &= \frac{P(x|M)\, P(M)}{P(x)} \propto P(x|M)\, P(M) \\ P(F|x) &= \frac{P(x|F)\, P(F)}{P(x)} \propto P(x|F)\, P(F)\,. \end{aligned}$$

[3] In common English usage, 'likelihood' is more-or-less a synonym for probability, perhaps with extra connotations of an event being hypothetical. In technical statistical usage, the likelihood is a property of an explanation of some data: a model or its parameters. The posterior probability of a parameter, model, or class, is proportional to its prior multiplied by its likelihood. Therefore saying 'likelihood of the data', while commonly-seen and acceptable informal English, conflicts with the 'correct' statistical usage: 'likelihood of the model (given the data)'.

x	$n_M(x)$	$P(x\|M)$	$n_F(x)$	$P(x\|F)$
1	0	0.00	0	0.00
2	0	0.00	0	0.00
3	0	0.00	0	0.00
4	0	0.00	2	0.02
5	1	0.01	7	0.07
6	2	0.02	8	0.08
7	2	0.02	10	0.10
8	2	0.02	14	0.14
9	7	0.07	21	0.21
10	8	0.08	19	0.19
11	14	0.14	10	0.10
12	22	0.22	4	0.04
13	19	0.19	2	0.02
14	11	0.11	1	0.01
15	6	0.06	1	0.01
16	2	0.02	1	0.01
17	2	0.02	0	0.00
18	1	0.01	0	0.00
19	1	0.01	0	0.00
20	0	0.00	0	0.00

Table 5.1: Example fish lengths for male and female, tabulated showing counts of each length per class, and relative frequencies used to estimate likelihoods for each class. (NB: These are different from the example data shown in section 5.1.)

In the case of this two-class problem:

$$\begin{aligned} P(x) &= P(x|M)\,P(M) + P(x|F)\,P(F) \\ P(M) + P(F) &= 1 \\ P(M|x) + P(F|x) &= 1\,. \end{aligned}$$

If we want to compare the posterior probabilities of M and F given the data we can take their ratio:

$$\begin{aligned} \frac{P(M|x)}{P(F|x)} &= \frac{P(x|M)\,P(M)/P(x)}{P(x|F)\,P(F)/P(x)} \\ &= \frac{P(x|M)\,P(M)}{P(x|F)\,P(F)}\,. \end{aligned} \tag{5.10}$$

In this case, if the ratio (Equation (5.10)) is greater than 1 then x is classified as M, if x is less than 1 then x is classified as F. As mentioned above, the denominator term $P(x)$ cancels.

Let's look at the same four test points as before. In this case let us assume that male and female fish have equal prior probabilities, that is $P(M) = P(F) = 1/2$.

1. $X = 5$
   $$\frac{P(M|X=5)}{P(F|X=5)} = \frac{P(X=5|M)\,P(M)}{P(X=5|F)\,P(F)} = \frac{0.01\cdot 0.5}{0.07\cdot 0.5} = 1/7$$
   Hence classify as $X=5$ as female (F).
2. $X = 16$
   $$\frac{P(M|X=16)}{P(F|X=16)} = \frac{P(X=16|M)\,P(M)}{P(X=16|F)\,P(F)} = \frac{0.02\cdot 0.5}{0.01\cdot 0.5} = 2$$
   Hence classify as $X=16$ as male (M).
3. $X = 9$
   $$\frac{P(M|X=9)}{P(F|X=9)} = \frac{P(X=9|M)\,P(M)}{P(X=9|F)\,P(F)} = \frac{0.07\cdot 0.5}{0.21\cdot 0.5} = 1/3$$
   Hence classify as $X=9$ as female (F).
4. $X = 12$
   $$\frac{P(M|X=12)}{P(F|X=12)} = \frac{P(X=12|M)\,P(M)}{P(X=12|F)\,P(F)} = \frac{0.22\cdot 0.5}{0.04\cdot 0.5} = 5.5$$
   Hence classify as $X=12$ as male (M).

Equal prior probabilities mean that we are equally likely to find a male or a female fish. If we believe that one sex is more prevalent than the other, then we can adjust the prior probability accordingly—and Bayes' theorem incorporates this information. Consider $X=11$; if the priors probabilities are equal (0.5), then we classify that value as male (M) since

$$\frac{P(M|X=11)}{P(F|X=11)} = \frac{P(X=11|M)\,P(M)}{P(X=11|F)\,P(F)} = \frac{0.14\cdot 0.5}{0.10\cdot 0.5} = 1.4$$

However, if we know there are twice as many females as males (i.e., $P(M) = 1/3$, $P(F) = 2/3$), then the ratio becomes

$$\frac{P(M|X=11)}{P(F|X=11)} = \frac{P(X=11|M)\,P(M)}{P(X=11|F)\,P(F)} = \frac{0.14\cdot /3}{0.10\cdot 2/3} = 0.7$$

and we classify the point as female (F).

The values we have been using for $P(x\,|\,M)$ and $P(x\,|\,F)$ are estimates, based on the relative frequencies—technically we refer to these as *maximum likelihood estimates*.[4]

Some questions. Assuming equal prior probabilities:

1. What is the value of $P(M \mid X=4)$?
2. What is the value of $P(F \mid X=18)$?
3. You observe data point $X=20$. To which class should it be assigned?

Comment on your answers, and any changes you might make to the probability estimates.

5.5 Optimisation problem

In section 5.3, we studied the basic idea of statistical pattern classification based on probabilities. To make the framework work properly, the probabilities need to be estimated from the training data as accurate as possible. To that end, we define a certain criterion to find what are supposed to be 'best'. This concept of optimality or optimisation plays an essential role in pattern recognition. For example, we will study parameter estimation of Gaussian distributions in Chapter 8, and training of neural networks in Chapters 11 and 12, both of which are defined as optimisation problems. We have already seen two examples of optimisation problem in Chapter 3. One is the mean squared error function for K-means clustering shown in Equation (3.1):

$$E = \frac{1}{N}\sum_{k=1}^{K}\sum_{n=1}^{N} z_{nk}\|\mathbf{x}_n - \mathbf{m}_k\|^2 \tag{5.11}$$

which we wanted to minimise with respect to the set of mean vectors $\{\mathbf{m}_k\}_1^K$. This can be formulated as the following optimisation problem:

$$\min_{\{\mathbf{m}_k\}_1^K} \frac{1}{N}\sum_{k=1}^{K}\sum_{n=1}^{N} z_{nk}\|\mathbf{x}_n - \mathbf{m}_k\|^2 \,. \tag{5.12}$$

Recall that the K-means clustering is an algorithm to find a local minimum solution of the optimisation problem. The second example is found in Equation (3.4), which is a formulation for PCA to find the first two principal components, $\mathbf{u}$ and $\mathbf{v}$, such that

$$\begin{aligned} &\max_{\mathbf{u},\mathbf{v}} \;\; \mathrm{Var}\,(y) + \mathrm{Var}\,(z) \\ &\text{subject to} \;\; \|\mathbf{u}\|=1, \|\mathbf{v}\|=1, \mathbf{u} \perp \mathbf{v}\,. \end{aligned} \tag{5.13}$$

Optimisation is a large field of study, which can be categorised in terms of types of variable (continuous or discrete) and types of constraint (unconstrained or constrained). The two examples above both fall into the category of continuous optimisation, but the former is a unconstrained optimisation problem, whereas the latter is a constrained one. In this section, we only consider continuous and unconstrained optimisation whose typical form [5] is given as

$$\min_{\mathbf{x}} f(\mathbf{x}) \tag{5.14}$$

[4] There are more sophisticated ways to deal with a model's unknown parameters, although these are beyond the scope of this course. For example, fully Bayesian methods consider all possible values of the parameters, weighted by the plausibility (posterior probability) of each setting.

[5] It does not matter whether we consider max or min, as $\max f(\mathbf{x})$ is equivalent to $\min -f(\mathbf{x})$.

where $\mathbf{x} \in \mathcal{R}^D$ and f is a smooth function such that $f : \mathcal{R}^D \to \mathcal{R}$. The function f to be optimised is called an *objective function*, and the solutions of the optimisation problem is called *optimal solutions*. It is not generally possible to find analytic (closed-form) solutions, and iterative optimisation algorithms are normally required as we have seen in the K-means clustering, but we here consider very simple cases where closed forms exist.

It is easy to see that the optimal solutions satisfy the following condition:

$$\frac{\partial f(\mathbf{x})}{\partial x_i} = 0, \text{ for } i = 1, \ldots, D \tag{5.15}$$

where $\frac{\partial f(\mathbf{x})}{\partial x_i}$ denotes the *partial derivative* of $f(\mathbf{x})$ with respect to the scalar variable x_i. This can be written in a vector form:

$$\left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_D}\right)^T = \mathbf{0} \tag{5.16}$$

where $\mathbf{0} = (0, \ldots, 0)^T$ is a zero vector in a D-dimensional vector space. The left hand side of this equation is called a *gradient* or a *gradient vector* of $f(\mathbf{x})$ and often denoted as $\nabla f(\mathbf{x})$ or $\mathrm{grad} f(\mathbf{x})$. Note that the condition above (i.e. $\nabla f(\mathbf{x}) = \mathbf{0}$) is not generally a sufficient condition for optimal solutions, but a necessary condition. However, the examples shown below consider a rather simple and specific case of quadratic functions in which the condition is sufficient to find optimal solutions.

5.5.1 Optimisation of a quadratic function of one variable

A simple yet meaningful example of optimisation problem would be the minimisation of the following quadratic function of a scalar variable x:

$$f(x) = ax^2 + bx + c \tag{5.17}$$

where $a > 0$, and b, c are arbitrary constants. By rewriting it into

$$ax^2 + bx + c = a\left(x + \frac{b}{2a}\right)^2 - \frac{b^2}{4a} + c \tag{5.18}$$

we can easily see that the solution is $x = -\frac{b}{2a}$. This can be also confirmed by using the necessary condition shown Equation (5.15) to get

$$\frac{\mathrm{d}f(x)}{\mathrm{d}x} = 2ax + b = 0 \tag{5.19}$$

which gives the same solution.

5.5.2 Optimisation of a quadratic function of two variables

We now consider the following function g of two variables, x and y :

$$g(x, y) = ax^2 + by^2 + cxy + dx + ey + f \tag{5.20}$$

where, for simplicity's sake, we assume $a > 0$, $b > 0$, $c^2 < 4ab$, and d, e, f are arbitrary constants. Taking partial derivatives with respect to x and y yields

$$\frac{\partial g}{\partial x} = 2ax + cy + d = 0 \tag{5.21}$$

$$\frac{\partial g}{\partial y} = 2by + cx + e = 0\,. \tag{5.22}$$

To find the optimal solution (x, y), we rewrite the above into a system of linear equations in matrix form:

$$\begin{pmatrix} 2a & c \\ c & 2b \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -d \\ -e \end{pmatrix}. \tag{5.23}$$

Thus

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2a & c \\ c & 2b \end{pmatrix}^{-1} \begin{pmatrix} -d \\ -e \end{pmatrix} \tag{5.24}$$

$$= \frac{1}{4ab - c^2} \begin{pmatrix} -2bd + ce \\ cd - 2ae \end{pmatrix}. \tag{5.25}$$

Question: Extend the above example to a quadratic function of three variables, x, y, and z. How about a case of n variables?

5.5.3 Line of best fit (Least squares fitting line)

Consider a set of N observations $\{\mathbf{p}_n\}_1^N$ in a 2D space, where $\mathbf{p}_n = (x_n, y_n)^T$, for which we would like to find the best fit line $y = ax + b$, where a and b are the parameters of the line.[6] As an objective function, we can consider the mean squared error between the predicted value $\hat{y}_i$ from x_i and the actual value y_i, so that the objective function E is given as:

$$E = \frac{1}{N} \sum_{n=1}^{N} (\hat{y}_n - y_n)^2 \tag{5.26}$$

where

$$\hat{y}_n = ax_n + b\,. \tag{5.27}$$

Thus, the optimisation problem is defined as $\min_{a,b} E$, i.e.,

$$\min_{a,b} \frac{1}{N} \sum_{n=1}^{N} (\hat{y}_n - y_n)^2\,. \tag{5.28}$$

Note that, as opposed to previous examples, this objective function E is not a function of x, but a function of a and b.

To solve the problem, we at first rewrite E by substituting Equation (5.27) into Equation (5.26).

$$E = \frac{1}{N} \sum_{n=1}^{N} (ax_n + b - y_n)^2\;. \tag{5.29}$$

The partial derivations of E with respect to a and b are given as

$$\frac{\partial E}{\partial a} = \frac{2}{N} \sum_{n=1}^{N} (ax_n + b - y_n)x_n = \frac{2}{N}\left(a\sum_{n=1}^{N} x_n^2 + b\sum_{n=1}^{N} x_n - \sum_{n=1}^{N} x_n y_n \right) \tag{5.30}$$

$$\frac{\partial E}{\partial b} = \frac{2}{N} \sum_{n=1}^{N} (ax_n + b - y_n) = \frac{2}{N}\left(a\sum_{n=1}^{N} x_n + b\sum_{n=1}^{N} 1 - \sum_{n=1}^{N} y_n \right). \tag{5.31}$$

[6] This is regarded as a *regression line* on y with x, whereas we can also consider a regression line on x with y, which is defined as $x = cy + d$.

By making the both partial derivatives, $\frac{\partial E}{\partial b}$ and $\frac{\partial E}{\partial b}$, equal to zero, and ignoring irrelevant coefficients, we get a system of linear equations in a and b:

$$a \sum_{n=1}^{N} x_n^2 + b \sum_{n=1}^{N} x_n = \sum_{n=1}^{N} x_n y_n \tag{5.32}$$

$$a \sum_{n=1}^{N} x_n + b \sum_{n=1}^{N} 1 = \sum_{n=1}^{N} y_n\,. \tag{5.33}$$

In matrix form,

$$\begin{pmatrix} \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n \\ \sum_{n=1}^{N} x_n & \sum_{n=1}^{N} 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{n=1}^{N} x_n y_n \\ \sum_{n=1}^{N} y_n \end{pmatrix}. \tag{5.34}$$

So, the solution of the optimisation problem is given as

$$\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{n=1}^{N} x_n^2 & \sum_{n=1}^{N} x_n \\ \sum_{n=1}^{N} x_n & \sum_{n=1}^{N} 1 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{n=1}^{N} x_n y_n \\ \sum_{n=1}^{N} y_n \end{pmatrix} \tag{5.35}$$

$$= \frac{1}{N\sum_{n=1}^{N} x_n^2 - (\sum_{n=1}^{N} x_n)^2} \begin{pmatrix} N\sum_{n=1}^{N} x_n y_n - (\sum_{n=1}^{N} x_n)(\sum_{n=1}^{N} y_n) \\ -(\sum_{n=1}^{N} x_n)(\sum_{n=1}^{N} x_n y_n) + (\sum_{n=1}^{N} x_n^2)(\sum_{n=1}^{N} y_n) \end{pmatrix}. \tag{5.36}$$

This can be simplified further, which is left as an exercise for the reader.

5.6 Summary

The main things covered in this chapter are:

- A simple example to demonstrate why we need to take uncertainty into account in pattern recognition
- A review of some of the main probability concepts that we will need: Independence; Conditional probability; The law of total probability; Bayes' theorem.
- The application of Bayes' theorem to statistical pattern recognition
- The basic idea of optimisation widely used in statistical pattern recognition

5.7 Reading

- Duda, Hart and Stork: Chapter 1, section 2.1
- Bishop: Sections 1.2.1, 1.2.2, 1.2.3

Exercises

1. Let X, Y, Z and $X_1, X_2, \ldots, X_N$ are random variables. Using the Bayes' theorem for two random variables, prove the following.

   (a) $P(X, Y, Z) = P(X|Y, Z)P(Y, Z)$

   (b) $P(X, Y|Z) = \dfrac{P(Y, Z|X)P(X)}{P(Z)}$

   (c) $P(Z|X, Y) = \dfrac{P(X|Y, Z)P(Z|Y)}{P(X|Y)}$

   (d) $P(X_1, X_2, \ldots, X_N) = P(X_1)\, P(X_2|X_1)\, P(X_3|X_2, X_1) \cdots P(X_N|X_{N-1}, \ldots, X_1)$

2. For the data set shown in Table 5.1, without using software tools, plot $P(x|F)$ and $P(x|M)$ by hand, and do the same for $P(F|x)$ and $P(M|x)$. (you may use a calculator if you want to)

3. In Section 5.5.3, we derived a formula for least squares fitting line, $y = ax + b$, which is also called a regression line (on y with x). Now, we consider a regression line on x with y, i.e., $x = cy + d$. Derive a formula to find the parameters, c and d.

4. Consider the following set of four observations, $\{\mathbf{p}_n\}_1^4$, where $\mathbf{p}_n = (x_n, y_n)^T$.

$$\mathbf{p}_1 = (1, 3)^T,\ \mathbf{p}_2 = (2, 0)^T,\ \mathbf{p}_3 = (10, 7)^T,\ \mathbf{p}_4 = (11, 4)^T$$

   (a) Find a regression line on y with x.

   (b) Find a regression line on x with y.