Lecture Outline

Reading: Chapter 1

- Review of probability theory (Section 1.2)

Note: Usually, we denote the probability that a random variable X takes on the value x as $\mathrm{P}(X = x)$ or $\Pr(X = x)$. In addition, probability density functions (pdfs) are denoted $p(X)$. In this text, however, we denote the probability that a random variable X takes on the value x as $p(X = x)$ and we also use the same notation, $p(X)$ to denote pdfs; $P(x)$ denotes a cumulative distribution.

1.2 Probability Theory

Discrete Random Variables

A key concept in the field of pattern recognition is that of uncertainty. It arises both through noise on measurements, as well as through the finite size of data sets. Probability theory provides a framework for the quantification and manipulation of uncertainty and forms one of the central foundations for pattern recognition. When combined with decision theory, it allows us to make optimal predictions given all information available to us, even though that information may be incomplete or ambiguous.

Consider the example shown in Fig. 1 involving two discrete random variables (RVs) X and Y. We shall suppose X can take on the values x_i for $1 \leq i \leq M$ and Y can take on the values y_j for $1 \leq j \leq L$. Consider N trials in which we sample both of the variables X and Y and let the number of trials in which $X = x_i$ and $Y = y_j$ be n_{ij}. Also let the number of trials in which $X = x_i$ be c_i and $Y = y_j$ be r_j.

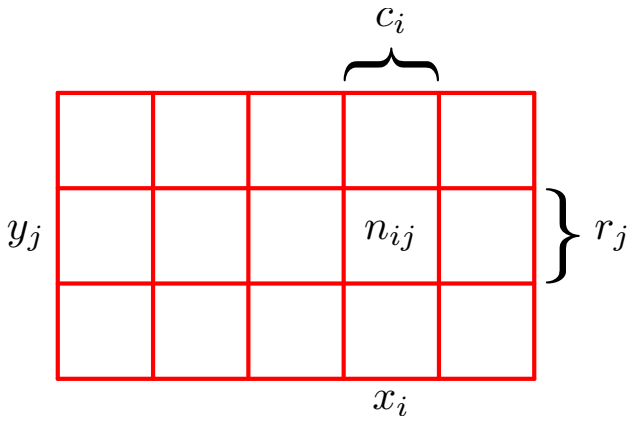

Figure 1: (Figure 1-10)

Joint Probability The probability that X will take on the value of x_i and Y will take on the value y_j is written $p(X = x_i, Y = y_j)$ and is called the *joint probability*. It is the number of points falling in cell i, j as a fraction of the total number of points

$$p(X = x_i, Y = y_j) \quad = \quad \frac{n_{ij}}{N}. \tag{1}$$

Marginal Probability Here we are implicitly considering the limit as $N \to \infty$. The probability that X will take on the value of x_i irrespective of the value of Y is written $p(X = x_i)$ and is given by the fraction

of the total number of points that fall in column i

$$p(X = x_i) = \frac{c_i}{N}. \tag{2}$$

Because the number of instances in column i is just the sum of the number of instances in each cell of that column we have

$$c_i = \sum_{j=1}^{L} n_{ij} \tag{3}$$

and therefore

$$p(X = x_i) = \sum_{j=1}^{L} p(X = x_i, Y = y_j) \tag{4}$$

which is the *sum rule* of probability. Note that $p(X = x_i)$ is sometimes called the *marginal probability* because it is obtained by marginalizing, or summing out, the other variables (Y in this case).

Conditional Probability If we consider only those instances for which $X = x_i$, then the fraction of instances for which $Y = y_j$, written $p(Y = y_j|X = x_i)$ and is called the *conditional probability* of $Y = y_j$ given $X = x_i$. It is obtained with the fraction of those points in column i that fall in cell i, j and given by

$$p(Y = y_j|X = x_i) = \frac{n_{ij}}{c_i}. \tag{5}$$

From the above, we can derive the *product rule* of probability

$$\begin{aligned} p(X = x_i, Y = y_j) &= \frac{n_{ij}}{N} \\ &= \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} \\ &= p(Y = y_j|X = x_i)p(X = x_i). \end{aligned} \tag{6}$$

In order to simplify the notation we will write $p(X)$ to denote the distribution over the RV X and $p(x_i)$ to denote the distribution evaluated for a particular value x_i. Thus our rules are compactly written

$$p(X) = \sum_{Y} p(X, Y) \tag{7}$$

$$p(X, Y) = p(Y|X)p(X) \tag{8}$$

Bayes' Theorem From the product rule and symmetry, $p(X, Y) = p(Y, X)$ we have Bayes' theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}. \tag{9}$$

Note that the denominator $p(X)$ can be written using the rules

$$p(X) = \sum_{Y} p(X|Y)p(Y). \tag{10}$$

We can provide an important interpretation of Bayes' theorem as follows. If we had been asked the value of Y before knowing the value of X, then the most complete information we have available is provided by $p(Y)$. We call this the *prior probability* because it is the probability available *before* observation of X. Once we are told the value of X, we can use Bayes' theorem to compute the probability of $p(Y|X)$, which we call

the *posterior* probability because it is the probability obtained *after* we have observed X. The *likelihood*, $p(X|Y)$ is how probable X is for different Y.

Finally, if

$$p(X,Y) = p(X)p(Y) \tag{11}$$

then X and Y are said to be *independent*. In this case from the product rule

$$p(Y|X) = p(Y) \tag{12}$$

and so the conditional distribution of Y given X is independent of the value of X.

Example: Consider two bins, red and blue. The red bin has 2 apples and 6 oranges and the blue bin has 3 apples and 1 orange as shown below. There are 12 pieces of fruit.

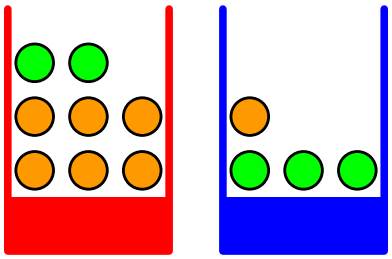

Figure 2: (Figure 1-9)

Let the RV X denote the fruit type and Y denote the bin color. A table describing the arrangement is shown below.

Table 1: Fruit in bins example.

		X	
		(a)pples	(o)ranges
Y	(r)ed	2	6
	(b)lue	3	1

Joint probabilities:

$$p(X=a,Y=r) = 2/12 = 1/6 \quad p(X=o,Y=r) = 6/12 = 1/2 \tag{13}$$

$$p(X=a,Y=b) = 3/12 = 1/4 \quad p(X=o,Y=b) = 1/12 \tag{14}$$

Note that when we sum the joint probabilities over the possible values of X and Y, we sum to 1.

Conditional probabilities (likelihoods): ("grabbing fruit out of the bin")

$$p(X=a|Y=r) = 2/8 = 1/4 \quad p(X=o|Y=r) = 6/8 = 3/4 \tag{15}$$

$$p(X=a|Y=b) = 3/4 \quad p(X=o|Y=b) = 1/4. \tag{16}$$

When we fix Y to a value and sum the conditional probabilities over the possible values of X, we sum to 1.

Prior probabilities: Suppose the prior probabilities for Y are given as $p(Y=r) = 2/5$ and $p(Y=b) = 3/5$. Note these priors have nothing to do with the data in the table–they are prior information regarding the *tendency* of Y.

Goal: We are interested in computing the *posterior* probability via Bayes' theorem

$$p(Y = r|X = o) = \frac{p(X = o|Y = r)p(Y = r)}{p(X = o)} \tag{17}$$

i.e. probability of a choosing a red bin given that the fruit we picked is an orange. We might be tempted to say that $p(Y = r|X = o) = 6/7$ but we are not taking into account the prior on Y thus we must use Bayes'.

Normalization factor (denominator): We might be tempted to say that $p(X = o) = 7/12$ but we are not taking into account that the fruit we get *depends* on which bin we picked to draw the fruit from which depends on the prior. The denominator in (17) is given by

$$\begin{aligned} p(X = o) &= p(X = o|Y = r)p(Y = r) + p(X = o|Y = b)p(Y = b) \\ &= \frac{3}{4} \cdot \frac{2}{5} + \frac{1}{4} \cdot \frac{3}{5} \\ &= \frac{9}{20}. \end{aligned} \tag{18}$$

Our *posterior* probability is thus

$$\begin{aligned} p(Y = r|X = o) &= \left(\frac{3}{4} \cdot \frac{2}{5}\right) / \frac{9}{20} \\ &= \frac{2}{3} \end{aligned} \tag{19}$$

which is viewed as the probability that $Y = r$ *given* the result (or observation or data) that $X = o$. Clearly the observation has changed our probability from $p(Y = r) = 2/5$ (prior to the observation) to $p(Y = r|X = o) = 2/3$ (posterior to the observation).

Example: Consider a "spam" filtering application where the random variable X denotes a word and Y denotes either that the email is ham (good) or spam (junk). Word occurences from training data, i.e. pre-classified email might produce the following (partial) table.

Table 2: Email data.

		X		
		'cash'	'medications'	'widget'
Y	(h)am	3	1	17
	(s)pam	9	12	6

Posterior probability: We would like to compute the *posterior* probability that a one-word email is ham given word content using Bayes theorem:

$$p(Y = \text{ham}|X = \text{cash}) = \frac{p(X = \text{cash}|Y = \text{ham})p(Y = \text{ham})}{p(X = \text{cash})}. \tag{20}$$

We might be tempted to say that $p(Y = \text{ham}|X = \text{cash}) = 3/(3 + 9)$ but we are not taking into account a prior on Y thus we must use Bayes'.

Conditional probabilities (likelihoods): From the training data, we have

$$p(X = \text{cash}|Y = \text{ham}) = 3/(3 + 1 + 17) = 1/7 \tag{21}$$

and

$$p(X = \text{cash}|Y = \text{spam}) = 9/(9 + 12 + 6) = 1/3 \tag{22}$$

Note that these conditional probabilities do not sum to 1.

Priors: Perhaps, we decide our *priors* based on email counts. For example, if 80% of the email we receive is spam, our priors would be

$$p(Y = \text{ham}) = 1/5 \quad \text{and} \quad p(Y = \text{spam}) = 4/5. \tag{23}$$

Normalization factor (denominator): Finally, we compute the normalization factor (word probabilities) as

$$\begin{aligned} p(X = \text{cash}) &= p(X = \text{cash}|Y = \text{ham})p(Y = \text{ham}) + p(X = \text{cash}|Y = \text{spam})p(Y = \text{spam}) \\ &= \frac{1}{7} \cdot \frac{1}{5} + \frac{1}{3} \cdot \frac{4}{5} \\ &= 0.2952 \end{aligned} \tag{24}$$

When we receive an email with the word 'cash' in it, we may compute (20) for both classes

$$\begin{aligned} p(Y = \text{ham}|X = \text{cash}) &= \frac{(1/7)(1/5)}{0.2952} \\ &= 0.0968 \end{aligned} \tag{25}$$

$$\begin{aligned} p(Y = \text{spam}|X = \text{cash}) &= \frac{(1/3)(4/5)}{0.2952} \\ &= 0.9032 \end{aligned} \tag{26}$$

and make a decision that the email is actually spam. In addition, for multi-word emails, we may compute the product of (20) using the word content in the email for each class (ham, spam) and base the decision on the larger of the two.