Tutorial 4: Introduction to statistical pattern recognition

1. The easiest way to do this is by plotting the points on a graph.

   Let the points be P_{A1} to P_{A4} and P_{B1} to P_{B4}. The distances from the new point P(2,3) to these points are:

$$\begin{aligned}
d(P, P_{A1}) &= \sqrt{(2-0)^2 + (2-2)^2} = 2 \\
d(P, P_{A2}) &= \sqrt{(2-0)^2 + (2-4)^2} = 2\sqrt{2} \\
d(P, P_{A3}) &= \sqrt{(2-1)^2 + (2-2)^2} = 1 \\
d(P, P_{A4}) &= \sqrt{(2-2)^2 + (2-3)^2} = 1 \\
d(P, P_{B1}) &= \sqrt{(2-2)^2 + (2-1)^2} = 1 \\
d(P, P_{B2}) &= \sqrt{(2-3)^2 + (2-1)^2} = \sqrt{2} \\
d(P, P_{B3}) &= \sqrt{(2-3)^2 + (2-3)^2} = \sqrt{2} \\
d(P, P_{B4}) &= \sqrt{(2-4)^2 + (2-4)^2} = 2\sqrt{2}
\end{aligned}$$

   k-nearest neighbour classification:

   $k = 3$: The 3 nearest points are P_{A3}, P_{A4} and P_{B1}, so P gets class A.

   $k = 5$: The 5 nearest points are P_{A3}, P_{A4}, P_{B1}, P_{B2} and P_{B3}, so P gets class B.

2. Let X be the variable which denotes mathematician m or engineer e. We know that, at the party:

$$\begin{aligned}
P(X=m) &= 0.2 \\
P(X=e) &= 0.8
\end{aligned}$$

   Let S be the variable denoting shoe-staring behaviour, $S=1$ denotes staring at shoes, $S=0$ not staring at shoes. Then

$$\begin{aligned}
P(S=1 \mid X=m) &= 0.6 \\
P(S=0 \mid X=m) &= 0.4 \\
P(S=1 \mid X=e) &= 0.1 \\
P(S=0 \mid X=e) &= 0.9
\end{aligned}$$

   We want to compute $P(X=m \mid S=1)$. Use Bayes' Theorem:

$$\begin{aligned}
P(X=m \mid S=1) &= \frac{P(S=1 \mid X=m)\,P(X=m)}{P(S=1)} \\
&= \frac{0.6 \cdot 0.2}{P(S=1)} \\
&= \frac{0.12}{P(S=1 \mid X=m)\,P(X=m) + P(S=1 \mid X=e)\,P(X=e)} \\
&= \frac{0.12}{0.6 \cdot 0.2 + 0.1 \cdot 0.8} \\
&= \frac{0.12}{0.20} = \frac{3}{5}.
\end{aligned}$$

   If you just wanted to know whether it was more probable that you were talking to a mathematician than an engineer, you could compute the *odds*:

$$\begin{aligned}
\frac{P(X=m \mid S=1)}{P(X=e \mid S=1)} &= \frac{P(S=1 \mid X=m)\,P(X=m)/P(S=1)}{P(S=1 \mid X=e)\,P(X=e)/P(S=1)} \\
&= \frac{P(S=1 \mid X=m)\,P(X=m)}{P(S=1 \mid X=e)\,P(X=e)} \\
&= \frac{0.6 \cdot 0.2}{0.1 \cdot 0.8} \\
&= \frac{0.12}{0.08} = \frac{3}{2}.
\end{aligned}$$

3. We have two random variables D (for disease) and T (for test).

   $D=1$ means the patient has the disease, $D=0$ means they do not.

   $T=1$ is a positive test, $T=0$ is a negative test,

   We can write the information in the question as:

$$\begin{aligned}
P(T=1 \mid D=1) &= 0.99 \\
P(T=0 \mid D=0) &= 0.95 \\
P(D=1) &= 0.01
\end{aligned}$$

   Since a there are only two outcomes having the disease (or not) and testing positive (or not) we can also write:

$$\begin{aligned}
P(T=0 \mid D=1) &= 1 - P(T=1 \mid D=1) = 0.01 \\
P(T=1 \mid D=0) &= 1 - P(T=0 \mid D=0) = 0.05 \\
P(D=0) &= 1 - P(D=1) = 0.99
\end{aligned}$$

   (a) *What percentage of subjects will test positive?*

   Use the law of total probability to calculate the number who test positive:

$$\begin{aligned}
P(T=1) &= P(T=1, D=1) + P(T=1, D=0) \\
&= P(T=1 \mid D=1)\,P(D=1) + P(T=1 \mid D=0)\,P(D=0) \\
&= 0.99 \times 0.01 + 0.05 \times 0.99 \\
&= 0.06 \times 0.99 \\
P(T=1) &= 0.0594
\end{aligned}$$

   (b) *Given that a subject tests positive, what is the posterior probability that they have the disease?*

Use Bayes' Theorem to find the posterior probability of having the disease given a positive test:

$$P(D=1 \mid T=1) = \frac{P(T=1 \mid D=1)\,P(D=1)}{P(T=1)}$$

$$= \frac{0.99 \times 0.01}{0.06 \times 0.99}$$

$$P(D=1 \mid T=1) = \frac{1}{6}$$

4. We can summarise the training data as follows, and estimate the likelihoods as relative frequencies:

R	T	D	$n(C=1)$	$P(\mathbf{x} \mid C=1)$	$n(C=0)$	$P(\mathbf{x} \mid C=0)$
0	0	0	1	0.05	6	0.3
0	0	1	1	0.05	3	0.15
0	1	0	2	0.1	4	0.2
0	1	1	1	0.05	1	0.05
1	0	0	3	0.15	2	0.1
1	0	1	3	0.15	3	0.15
1	1	0	4	0.2	0	0
1	1	1	5	0.25	1	0.05

Classify the test data by computing the posterior probabilities. If $P(C=1 \mid X) > 0.5$, classify as $C=1$, else classify as $C=0$ (since it is a two class problem).

$$P(C=1 \mid X) = \frac{P(X \mid C=1)\,P(C=1)}{P(X \mid C=1)\,P(C=1) + P(X \mid C=0)\,P(C=0)}$$

$$P(C=1 \mid \mathbf{x}_1) = \frac{0.25 \cdot 0.25}{0.25 \cdot 0.25 + 0.05 \cdot 0.75}$$

$$= \frac{0.0625}{0.1} = \frac{5}{8}$$

$$P(C=1 \mid \mathbf{x}_2) = \frac{0.15 \cdot 0.25}{0.15 \cdot 0.25 + 0.1 \cdot 0.75}$$

$$= \frac{0.0375}{0.1125} = \frac{1}{3}$$

$$P(C=1 \mid \mathbf{x}_3) = \frac{0.1 \cdot 0.25}{0.1 \cdot 0.25 + 0.2 \cdot 0.75}$$

$$= \frac{0.025}{0.175} = \frac{1}{7}$$

So classify $\mathbf{x}_1$ as $C=1$, and $\mathbf{x}_2$ and $\mathbf{x}_3$ as $C=0$.

Although the training data happens to balanced (and we don't know how it was collected), it is the case that that the priors are not equal, and this makes a difference (e.g., in the classification of $\mathbf{x}_2$).

Directly estimating the likelihood is possible if we have a limited number of feature vectors $2^3 = 8$ in this case. But if the feature dimension increases (e.g., if we have 10 binary dimensions, then we have $2^{10} = 1024$ possible feature vectors) or the number of possible values increases (e.g., if we our 3 attributes can each take 5 possible values we have $5^3 = 125$ possible feature vectors). The more possible feature vectors we have the more training data we need to estimate $P(X \mid C)$. The example of this question already has problems with limited data: we estimate $P(X = (1\ 1\ 0) \mid C = 0) = 0$, which is surely an underestimate—just because we do not observe something in a few tens of examples does not mean it will never happen.

One approximation to cope with the 'curse of dimensionality' is Naive Bayes, which treats each input dimension independently, i.e. in this case it assumes $P(R, T, D) = P(R)\,P(T)\,P(D)$—fewer probabilities to estimate, but a big approximation. This is the topic of this week's learning lectures.