Tutorial 2: Introduction to statistical pattern recognition

1. Given a two dimensional space with the following dataset:

   Class A:  (0, 2),  (0, 4),  (1, 2),  (2, 3)
   Class B:  (2, 1),  (3, 1),  (3, 3),  (4, 4)

   Classify a new point (2, 2) using $k$-nearest neighbour classification using Euclidean distances and $k = 3$ and $k = 5$.

2. 60% of mathematicians stare at your shoes when they meet you, but only 10% of engineers do. You are at an exciting party composed entirely of mathematicians and engineers. 80% of the people there are engineers. You meet someone who stares at your shoes. What is the probability that they are a mathematician?

3. A screening test is devised for a disease. It seems that the test is very accurate: 99% of people with the disease test positive; 95% of people who do not have the disease test negative. Of those who are given the test, 1% actually have the disease.

   (a) What percentage of subjects will test positive?

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

4. Consider a fictitious medical condition $C$, which is either present ($C = 1$) or absent ($C = 0$) in a subject. The only information we have about a subject is whether they have a rash ($R = 1$), have a temperature ($T = 1$), or are dizzy ($D = 1$). Thus we have a 3-dimensional feature vector, $(R, T, D)$. If we have the following information about a subject: $R = 1$, $T = 0$, $D = 1$, then the feature vector is $X = (1, 0, 1)$.

   Training data are available from 40 subjects, shown in figure 1 (overleaf). Using this training data, estimate the likelihoods:

   
   $P(X = (0, 0, 0) \mid C = 1), \ldots, P(X = (1, 1, 1) \mid C = 1), \ldots, P(X = (0, 0, 0) \mid C = 0), \ldots, P(X = (1, 1, 1) \mid C = 0)$.

   The following test data are observed:

   $x_1 = (1, 1, 1), \quad x_2 = (1, 0, 0), \quad x_3 = (0, 1, 0)$.

   It is known that the prior probability of the condition is $P(C = 1) = 0.25$. To which class should each test vector be classified?

   Comment on this approach to classification if we had a situation with a 10-dimensional feature vector, or if we have a situation where each input dimension has 5 possible values rather than 2.
5. This is an extension of the line of best fit discussed in Section 5.5.3 in Lecture Note 5 to a 3D case. Consider a set of \( N \) observations \( \{p_n\}_{n=1}^N \) in a 3D space, where \( p_n = (x_n, y_n, z_n)^T \), for which we would like to find the best fit plane \( z = ax + by + c \). Derive the system of linear equations in \( a, b, \) and \( c \). (NB: It is more general to define a plane as \( ax + by + cz + d = 0 \), but we here consider a simpler version.)