1. **K-nearest neighbour (KNN) classifier. (10 points)**
   a. Name one advantage and one limitation of KNN (4 points).

   Advantage: flexible, nonlinear decision boundary
   Limitation: memory footprint, computational cost

   b. In the figure below, there are 4 samples, two of them are from positive class (circles) and two from negative class (triangles). Draw the class boundary of KNN for $k = 1$. (6 points)

2. **Feature normalization and selection. (15 points)**
   Given a data matrix $X \in \mathbb{R}^{n \times d}$, with $n$ samples and $d$ features, and label vector $Y \in \mathbb{R}^{n \times 1}$ for regression task.

   a. Provide two ways to normalize the features in case their numerical ranges vary too much compared with each other. (4 points)

   Method 1. Deduct each feature by the mean and divided by the standard deviation
   Method 2. Deduct each feature by the min and divide by the (max – min) of that feature.

   b. How to identify the top-$k$ most useful features? Write down any formula needed (6 points)

   Use correlation coefficient between the feature ($x$) and the label ($Y$). Those with largest absolute values will be the most useful features.
3. **Conditional probability and Information gain. (20 points)**
   a. Write down Bayesian conditional probability $P(X|Y)$ for random variable $X, Y$; (**3 points**)
   
   $$ P(X|Y) = \frac{P(X,Y)}{P(Y)} $$

   b. Given the table between two random categorical variables, “Temperature” ($X = \text{High, Median, Low}$) and “Rain” ($Y = \text{Yes, No}$). Compute the following
   
   \begin{itemize}
   \item $P(X = \text{High})$ (**2 points**) \( \frac{2}{10} \)
   \item $P(X=\text{Low, Y = No})$ (**2 points**) \( \frac{3}{10} \)
   \item $P(X = \text{High}|Y = \text{Yes})$ (**3 points**) \( \frac{2}{5} \)
   \end{itemize}

   c. What is the entropy of the random variable $Y$? (**3 points**)
   
   \( 2 \times 0.5 \times \log_2(0.5) \rightarrow 1 \)

   d. What is the information gain on $Y$ after partitioning it by variable $X$? (**7 points**)

   After partitioning samples based on $X$, we have three groups:
   
   - Group 1: $X$ is high, $Y$ is pure (2 yes) so the entropy is 0;
   - Group 2: $X$ is Median and $Y$ is mixed (2 yes 2 no), so the entropy is 1
   - Group 3: $X$ is low and $Y$ is pure (2 no), so the entropy is 0;
   
   So the weighted entropy is
   \[
   \frac{2}{10} \times 0 + \frac{4}{10} \times 1 + \frac{2}{10} \times 0 = 0.4
   \]
   
   The information gain is $1 - 0.4 = 0.6$

4. **Logistic regression. (30 points)**
   a. Explain what is maximum a posterior estimation, given the data $D$ and your model/hypothesis $h$. Clearly name every term in your formula, and how is it related to the maximum likelihood estimation? (**6 points**)

\[
 r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}
\]
Posterior: P(h|D); Likelihood: P(D|h); Prior: P(h).

MAP estimation means that we want to choose the hypothesis h that maximizes the posterior probability p(h|D).

Maximum likelihood estimation assumes that the prior p(h) is uniform and so we want to only maximize the likelihood function p(D|h).

b. Suppose one wants to learn a model with parameter w (vector), and what kind of prior distribution p(w) can you think of that will encourage a sparse model when maximizing the logarithm of the posterior probability? (4 points)

Let \( p(w) = \frac{1}{\pi |w|^2} \), then \( \log(p(w)) = -\frac{1}{2} |w|^2 \), and so maximizing logarithm of the posterior \( p(h|D) \) requires maximizing \( \log(p(D|h)) + \log(p(w)) = \log(p(D|h)) - \frac{1}{2} |w|^2 \), which is equivalent to minimizing \( \frac{1}{2} |w|^2 \) and this will encourage sparsity.

c. In logistic regression with two classes (1 and 0)
   i. How is the odds ratio is approximated? (3 points)
   ii. How to compute the probability of a sample belonging to the positive class (1), namely \( p(1|x) \) where \( x \) is a m-dimensional sample point \( x = [x_1, x_2, ..., x_m] \) (5 points)
   iii. How to derive the (log)-likelihood function for logistic regression given the samples and their labels \( \{x_i, y_i\}_{i=1}^{n} \) where \( i \) is sample index and \( n \) is sample size? (7 points)

In logistic regression one uses a linear function to approximate the logarithm of the odds ratio, namely

\[
\frac{P(1|x_1, x_2, ..., x_m)}{P(0|x_1, x_2, ..., x_m)} \cdot \frac{1}{P(1|x_1, x_2, ..., x_m)} \cdot (w_1 x_1 + w_2 x_2 + ... + w_m x_m \xi b)
\]

The probability that a sample belongs to positive class is

\[
P(1|x_1, x_2, ..., x_m) = \frac{1}{1 + e^{-(w_1 x_1 + w_2 x_2 + ... + w_m x_m + \xi b)}}
\]

The overall log-likelihood for the samples and the labels is

\[
\log \text{-likelihood} = \log(p(x_i))
\]

5. Outlier detection. In unsupervised setting, outliers are those samples that are far away from others. Given a sample set \( \{x_i\} \) for \( i = 1, 2, ..., n \). If we assume that the samples are drawn from a univariate Gaussian distribution, how to detect the outliers? (10 points)

Assume that the Gaussian distribution is \( f(x) = \frac{1}{2\pi h} \exp \left( -\frac{||x-\mu||^2}{2h^2} \right) \)

Then the mean and standard deviation of the Gaussian distribution can be estimated as
\[ \mu = \frac{1}{n} \sum_{i=1}^{n} x_i \]
\[ h^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 \]

After estimating the parameters of this Gaussian distribution, we can plug each sample \( x_i \) into \( f(x) \) and examine the probability for each sample; those with a small probability \( f(x_i) \) will be outliers. One can also simply calculate the distance between \( x_i \) and \( \mu \) and if \( \|x_i - \mu\|_2 > 3h \) then the sample \( x_i \) falls out the 95% probability mass and can be deemed as outlier.

6. **Bayesian Conditional Probability. (15 points)**

   Consider diagnosis of Cancer based on the value of a certain biomarker test with numerical range [0,60]. In the Figure below, we plot the histogram of the historical test results of 100 human subjects. Some subjects are healthy (blue) and some are cancerous (red). The doctor will release a positive report if the test value > 45, or negative if the test value <45. Let \( C \) be cancerous and \( \sim C \) be healthy; Let \( P \) denotes positive result, and \( N \) denotes a negative result.

   a. Compute the prior distribution of \( P(C) \) and \( p(\sim C) \), as well as the conditional probabilities \( p(P|C) \), \( p(N|C) \), \( p(P|\sim C) \), \( p(N|\sim C) \). (6 points)

   \[
   \begin{align*}
   P(C) &= 0.20; \\
   P(\sim C) &= 0.80; \\
   p(P|C) &= 0.75; \\
   p(N|C) &= 0.25; \\
   p(P|\sim C) &= 1/16; \\
   p(N|\sim C) &= 15/16
   \end{align*}
   \]

   b. Compute the joint probabilities \( p(P,C) \), \( p(N,C) \), \( p(P,\sim C) \), \( p(N,\sim C) \). (4 points)

   \[
   \begin{align*}
   p(P,C) &= 0.15; \\
   p(N,C) &= 0.05; \\
   p(P,\sim C) &= 0.05; \\
   p(N,\sim C) &= 0.75
   \end{align*}
   \]

   c. A new subject gets a test and the result is 50. What is the probability of the subject to be truly cancerous? (5 points)

   \[
   P(C|P) = \frac{p(C,P)}{p(P)} = \frac{p(C,P)}{(p(C,P) + p(\sim C,P))} = \frac{0.15}{0.15 + 0.05} = 75\%
   \]
Suppose we have a number of 1-dimensional samples drawn from two classes, as shown in the figure below; the decision function is defined as \( f(x) = \begin{cases} 1, & \text{if } x > c \\ -1, & \text{if } x \leq c \end{cases} \)

Obviously, the performance of classifier depends on the choice of the threshold \( c \).

![Precision-Recall Curve](image)

1. Tune the value of \( c \) and for each value, mark the precision-recall on the right figure. You can follow the suggested \( c \) values in the axis (marked as stars) (7 points)

2. What is the best choice of \( c \), and why? (3 points)

The best \( c \) corresponds to the point on the precision-recall curve that is farthest away from the off-diagonal line, namely \( c = \) [Value]