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

   ![KNN Class Boundary](image)

2. **Feature normalization and selection. (10 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)**
   
   b. How to identify the top-$k$ most useful features? Write down any formula needed **(6 points)**.
3. **Conditional probability and Information gain. (20 points)**
   a. Write down Bayesian conditional probability $P(X|Y)$ for random variable $X$, $Y$; **(3 points)**

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

   c. What is the entropy of the random variable $Y$? **(3 points)**
      - $\log_2 0 = \phi$
      - $\log_2 0.5 = -1$
      - $\log_2 1 = \phi$

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

4. **Logistic regression and regularization. (25 points)**
   a. Explain what is maximum a posterior estimation (MAP), given the data $D$ and your model/hypothesis $h$. Clearly name every term in your formula. How is MAP different from maximum likelihood estimation (MLE)? **(6 points)**

   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)**

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

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.

![Histogram of test results](image)

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)

\[
\begin{align*}
P(C) & =; \\
P(\sim C) & =; \\
p(P|C) & =; \\
p(N|C) & =; \\
p(P|\sim C) & =; \\
p(N|\sim C) & =; 
\end{align*}
\]

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

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

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)
7. **Evaluation metric (Precision and Recalls) (10 points)**

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 \(c\).

\[\begin{array}{cccccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\
\end{array}\]

\(\bigcirc\) Positive Class (1)

\(-\) Negative Class (-1)

(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)