Rehearsal Exercise

1. The Bias and Variance Dilemma
   a. Explain what are the Bias and Variance of a model, and the tradeoff between them (you can use a schematic draw for illustration)
   b. What is the effect of ensemble methods on the Bias and Variance
      i. Give at least one specific example for an ensemble method that reduces either the Bias or the Variance (or both)

2. LDA can be considered as a classifier
   a. Define the classification model, and explain the primal probabilistic assumption, the logit transformation, and the decision boundary implied by the model
   b. Derive the LDA solution for a two-class setting
   c. Is LDA a generative or a discriminative model?
   d. Give a 2-dim example where the LDA solution will be better than an ensemble of linear classifiers constructed using 1-vs-all scheme

3. Given a set of \( n \) observations \( x_1, ..., x_n \)
   a. Assume that the observations are Gaussian distributed, i.e.
      \[
      \Pr(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right],
      \]
      Calculate the Maximum Likelihood Estimator (MLE) for the parameters \( \mu, \sigma^2 \)
   b. Assume that the observations were generated by \( k \) Gaussian distributions, all of which are with a unit variance, i.e. \( \sigma^2 = 1 \)
      i. Define the mixture model
      ii. Use the Expectation Maximization (EM) algorithm to estimate the parameters of the mixture model

4. Ensemble Learning
   a. Why use ensemble based systems?
   b. What are the key properties the base learner should possess so that the ensemble performance will surpass the base learner performances?
   c. Describe the leading schemes for combining base learners
   d. What is common and what is different between Bagging and Random Forests?

5. Bayes Learning
   a. Explain the Bayes formula and show how it assists in solving a classification problem.
   b. In what sense a Bayes decision rule is optimal (note the question is about any Bayes decision rule and not just Naive Bayes Classifier).
   c. What is the difference between Maximum-A-Posteriori (MAP) hypothesis and Maximum Likelihood (ML) hypothesis? In what conditions can you use ML and in what MAP?

6. Linear Methods for Classification
a. What are the two classes of methods to model linear decision boundaries? Discuss the differences, and provide examples of at least one algorithm from each class.

b. Define the Perceptron Criterion, i.e. the error function of the Perceptron algorithm, and derive the Perceptron learning rule.

c. Define the error function of the Widrow-Hoff algorithm, and derive the Widrow-Hoff learning rule.

7. Boosting
A decision stump is a one-level decision tree. That is, it is a decision tree with one node (the root) which is immediately connected to the terminal nodes (its leaves). A decision stump makes a prediction based on the value of just a single input feature. Decision stump is a very popular weak learner which is often used for boosting. Josef, a student in the course, suggested an improvement – Decision stamp 2.0, which is a two-level decision tree. Josef believes that Decision stamp 2.0 is a much better weak learner, and he argues that with Decision stamp 2.0 training will be much faster and with a smaller rounds of boosting (namely the resulted committee will be smaller). Do you agree? Discuss the pros and cons vs. using the original (one-level) decision stamp.

8. Given the samples below in \( \mathbb{R}^2 \), for each of the following classifiers, describe whether it can achieve zero classification error. If it can achieve zero classification error, draw a possible decision boundary – Linear SVM, Decision Tree, Gaussian Naïve Bayes and KNN (specify the value of \( k \)).

9. What is the VC dimension of a generalized Parabola in \( \mathbb{R}^2 \) (parabola that may be tilted on the plane)?