Problems 1 and 5 were graded by Amin Sorkhei, Problems 2 and 3 by Johannes Verwijnen and Problem 4 by Jyrki Kivinen.

1. [*12 points*]

   (a) *Gini index* and *Entropy* are **impurity measures** which can be used in order to measure the impurity of subset D of training data set. These measure are defined based on the **following formulas**, where k stands for the number of classes and p_i is the fraction of examples of class i among all examples in D:

   $$Entropy(D) = -\sum_{i=1}^{k} p_i \log(p_i)$$

   $$Gini(D) = 1 - \sum_{i=1}^{k} p_i^2$$

   Directions: In order to receive full mark for this question, one needs to mention that the above measurements are mainly impurity measurements (1 points) and also include the formulas regarding each measurement (1.5×2 points).

   (b) *one-versus-one* and *all-versus-all* are methods which can be used in order to use a **binary classifier** to do **multi-class classification**. In *one-versus-one* method, **for each pair of distinct classes** as classifier is trained based on the corresponding data points in the training set. In order to classify,a data point is classified based on **all classifiers** and the label with the **most votes** is selected as the class label for that data point.
   In *one-versus-rest* method, **for each class label** i a classifier is trained where the corresponding data points are considered with (+) label and the rest is considered with (-) label and a wight vector w_i is returned. In order to **classify** a data point x, one can simply take the class label which maximizes $w_i^T x$.
   Directions: In order to receive full mark for this question, one needs to mention the training process (1×2 points) and the classification process (1×2 points) for each method.

   (c) *single-linkage* and *complete-linkage* are among linkage functions in order to generalize the notion of dissimilarity between two points ($Dis(x, y)$) to **two sets of points** ($Dis(A, B)$).
   single-linkage considers the closest pair of data point in clusters.

   $$L_{single}(A, B) = \min_{x \in A, y \in B} Dis(x, y)$$

   complete-linkage considers the maximum dissimilarity between data points in clusters.

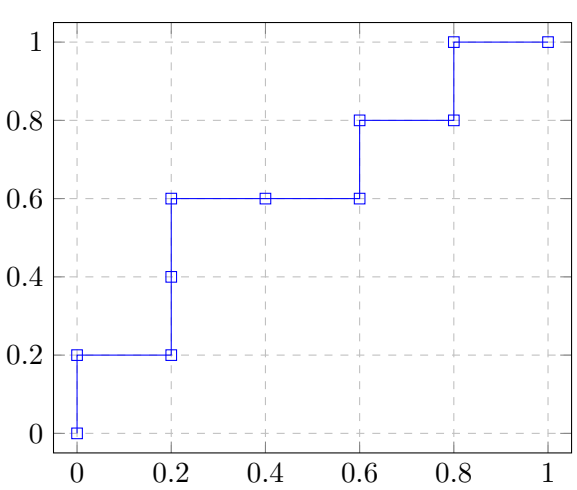
   $$L_{complete}(A, B) = \max_{x \in A, y \in B} Dis(x, y)$$

   Directions: In order to receive full mark for this question, one needs to define the linkage functions (2×1 points) and also include the formula and example for

each case (1×2 points).

2. [*12 points*]

   (a) The procedure for drawing a ROC curve is given on slides 106-110, book pages 63-67. One first needs to order the scores given by $\hat{s}(x)$ into decreasing order. One then starts from the point $(0, 0)$ and draws a horizontal or vertical line segment for each data point depending on whether its true label is positive or negative.



   The question explicitly asks one to *draw* the ROC curve. 8 points were given for a correct ROC curve, 2 points for at least ordering the scores, minus points for smaller problems in the curve.

   (b) The ranking accuracy of a classifier is defined on slide 105 (via ranking error) and book page 64 as

$$rank - acc = \frac{\sum_{x \in Te^+, x' \in Te^-} \left(I\left[\hat{s}(x) > \hat{s}(x')\right] + \frac{1}{2} I\left[\hat{s}(x) = \hat{s}(x')\right] \right)}{Pos \cdot Neg}$$

   It can be determined from the ROC curve by calculating the AUC (Area Under ROC Curve). In the example, the AUC, and thus ranking accuracy is $\frac{16}{25}$.

   2 points for the definition, one point each for how to determine it from a ROC and the value for the example.

3. [*12 points*] We're given $P(X_1)$, $P(X_2|Y)$ and $P(Y)$ and asked for $P(Y = +1|X)$ for all X. We know that $P(X) = P(X_1, X_2) = P(X_1)P(X_2)$. Using Bayes' rule we get.

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

   For the different values of X we get proportional probabilities

$$
\begin{aligned}
P(Y=1|X=(1,1)) &\propto \frac{1}{2}\times\frac{3}{4}\times\frac{2}{3}=\frac{1}{4}\\
P(Y=1|X=(1,-1)) &\propto \frac{1}{2}\times\frac{1}{4}\times\frac{2}{3}=\frac{1}{12}\\
P(Y=1|X=(-1,1)) &\propto \frac{1}{2}\times\frac{1}{4}\times\frac{2}{3}=\frac{1}{12}\\
P(Y=1|X=(-1,-1)) &\propto \frac{1}{2}\times\frac{3}{4}\times\frac{2}{3}=\frac{1}{4}\\
P(Y=-1|X=(1,1)) &\propto \frac{1}{2}\times\frac{1}{4}\times\frac{1}{3}=\frac{1}{24}\\
P(Y=-1|X=(1,-1)) &\propto \frac{1}{2}\times\frac{3}{4}\times\frac{1}{3}=\frac{1}{8}\\
P(Y=-1|X=(-1,1)) &\propto \frac{1}{2}\times\frac{3}{4}\times\frac{1}{3}=\frac{1}{8}\\
P(Y=-1|X=(-1,-1)) &\propto \frac{1}{2}\times\frac{1}{4}\times\frac{1}{3}=\frac{1}{24}
\end{aligned}
$$

These results seem to follow the intuition to predict $Y=1$ when $X_1=X_2$ and $Y=-1$ otherwise. The question asked us to provide the posterior probabilities and for that we have to normalize per $P(X)$, which we get by the Law of Total Probability $P(X)=\sum_Y P(X|Y)P(Y)$, thus

$$
\begin{aligned}
P(Y=1|X=(1,1)) &= \frac{\frac{1}{4}}{\frac{1}{4}+\frac{1}{24}}=\frac{\frac{6}{24}}{\frac{7}{24}}=\frac{6}{7}\\
P(Y=1|X=(1,-1)) &= \frac{\frac{1}{12}}{\frac{1}{12}+\frac{1}{8}}=\frac{\frac{2}{24}}{\frac{5}{24}}=\frac{2}{5}\\
P(Y=1|X=(-1,1)) &= \frac{\frac{1}{12}}{\frac{1}{12}+\frac{1}{8}}=\frac{\frac{2}{24}}{\frac{5}{24}}=\frac{2}{5}\\
P(Y=1|X=(-1,-1)) &= \frac{\frac{1}{4}}{\frac{1}{4}+\frac{1}{24}}=\frac{\frac{6}{24}}{\frac{7}{24}}=\frac{6}{7}
\end{aligned}
$$

The Bayes-optimal classifier for this distribution is thus $Y=1$ if $X=(1,1)$ or $X=(-1,-1)$ and $Y=-1$ otherwise.
The Bayes Error is the sum of probabilities when making an error, which we can sum up from the proportional distributions as $\frac{1}{12}+\frac{1}{12}+\frac{1}{24}+\frac{1}{24}=\frac{1}{4}$.
8 points were awarded for the posterior probabilities, single points were subtracted for (very common) minor arithmetic mistakes. 2 points were subtracted for unnormalized probabilities.
2 points were awarded for the correct Bayes-optimal classifier and Bayes Error each.

4. Explaining the terms *generalisation* and *overfitting* gave **4 points**. Most students scored very well in this part of the problem. Sometimes a point was deducted for answers that were too vague or otherwise poorly expressed. It was not required that the answer

should use or explain terms "bias" and "variance," but on the other hand just using the terms without any explanation of their meaning did not give any points.

Another **4 points** was given for explaining the procedure. A complete answer should explain withholding part of the available training data as a separate test set, and the cross-validation procedure. It should also mention the combination of first using cross-validation (or a validation set) to choose the correct model complexity, and the using a test set to evaluate the generalisation of the final result. Most students were able to explain the individual techniques well, but explanations of the whole process were less clear. In particular, it should be noticed that cross-validation is not in itself a way to avoid overfitting. It is a tool for evaluating generalisation, which enables us to choose parameters of our algorithm etc. in order to optimise generalisation and thus avoid overfitting.

The remaining **4 points** were given for two examples, two points for each example. Full score required explaining both the model-specific technique and its connection to the general procedure. For example, two points was given for giving the formula for ridge regression and mentioning that the regularisation parameter λ is chosen by cross-validation, or by explaining the reduced error pruning process and mentioning that it must use a separate pruning set.

5. [*12 points*]

   (a) *K-means* is an **unsupervised** method in order to **cluster** the input data. More precisely, *K-means* receives **data set** and **number of clusters** (initial centroids) as input and returns **final centroids** and **assignment mappings** as output. Intuitively, members of final clusters must be similar to each other and must be the least dissimilar to corresponding centroids as compared to other points. More precisely, one can interpret the results using the classification-like method, F-measure and Silhouettes coefficients.

   (b) *K-means* is mainly composed of four steps, three of which occur repeatedly until convergence.

   - Randomly choose initial centroids $\mu_1, \mu_2, ..., \mu_k$.
   - Repeat the following steps until convergence
     - Determine clusters: for $i = 1 \dots n$, let $j(i) \longleftarrow \arg\min_j ||x_i - \mu_j||_2^2$
     - Update cluster assignments: for $j = 1 \dots k$, let $D_j \longleftarrow \{x_i | j(i) = j\}$
     - Update Centroids: for $j = 1 \dots k$, let $\mu_j \longleftarrow \frac{1}{|D_j|} \sum_{x \in D_j} x_i$

   (c) *K-means* tries to minimize the following objective cost function:

$$cost = \sum_{j=1}^{k} \sum_{x \in D_j} ||x - \mu_j||_2^2 = \sum_{i=1}^{n} ||x_i - \mu_{j(i)}||_2^2$$

   It is guaranteed that *K-means* minimizes this function after a finite number of steps, while, it not at all guaranteed that the global minimum is found.
   Directions:

i. In order to receive full mark for the first part, one needs to mention that *K-means* is used to cluster the data (and preferably is an unsupervised method) (1 point) and the input (1 point) and the output(1 point) and and intuitive way to interpret the results (1 point).

ii. In order to receive full mark for part b, one needs to carefully describe all the mentioned steps (1 point) each.

iii. In order to receive full mark for part c, one needs to include any of the formulas mentioned above (2 points), mention that the cost function is minimized (1 point) but the global minimum is not guaranteed to be found (1 point).