1. You have 50 minutes to finish the exam.

2. There are 7 pages in this exam (including cover page), please write down your initials on top of EVERY page.

3. If you use the back of the page please indicate on the front of the page so I won’t miss it.

Prob 5 is turned into a bonus prob.

Old Total - prob 5 \[ \frac{64}{64} \] + prob 5 score.

<table>
<thead>
<tr>
<th>Page</th>
<th>Max</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>74</td>
<td></td>
</tr>
</tbody>
</table>
1. Short questions.

- (6 pts) We have two identical bags A and B. A holds 2 red and 4 green marbles. B holds 4 red and 4 green marbles. Randomly pick a bag, and draw a marble, what is the probability for green?

\[ P(G) = P(G|A)P(A) + P(G|B)P(B) = \frac{4}{6} \times \frac{1}{2} + \frac{4}{8} \times \frac{1}{2} = \frac{7}{12} \]

- (4 pts) K-means clustering is sensitive to initialization. One approach to address this is to try multiple random restarts, each with a different initialization. This gives us multiple different clusterings, how should we choose among them?

Choose the one with the lowest objective. (i.e. SSE).

value.

- (7 pts) Consider a data set with 5 binary features plus one categorical feature with 10 categories. Using the information gain criterion to build a decision tree for this data, we'll more likely select the categorical feature, which will create 10 branches. What is the reason for this bias? Suggest a strategy to avoid this issue.

Ten branches likely lead to pure nodes after split compared to two branches.

Thus more branches leads to more information gain.

Strategy: create a binary splits by comparing against one possible category.
- (2 pts each) Please indicate for each action below whether it increases, decreases, or does not impact overfitting for a classification problem.
  * For KNN, change $k$ from 1 to 5
    
    - Decrease
  
  * Decrease the number of training examples
  
    - Increase
  
  * Change from a linear kernel to a 3rd order polynomial kernel in SVM.
  
    - Decrease
  
  * Change from a Bayes classifier to a Naive Bayes classifier
  
    - Decrease
  
  * Apply PCA to reduce the feature dimension by half.
  
    - Decrease

- (6 points) For hierarchical agglomerative clustering, what is the difference between single-link and complete-link? Consider the data set below: cluster 1 forms an 's', and the other points belong to cluster 2, which of the two linkage methods will work well?

\[
\text{Single: } D(C_1, C_2) = \min_{x \in C_1, y \in C_2} D(x, y)
\]

\[
\text{Complete: } D(C_1, C_2) = \max_{x \in C_1, y \in C_2} D(x, y)
\]

Single link is better for this because it can find long straggly clusters.
2. (10 pts) True or false.

a. (2 pts) Online Perceptron updates the current \( w \) vector when it misclassifies a training example \((x_i, y_i)\) using update rule \( w \leftarrow w + y_i x_i \). After the update, the new \( w \) is guaranteed to correctly classify \((x_i, y_i)\).

   \[ \text{False. It corrects in the right direction, but does not guarantee all the way.} \]

b. (2 pts) When applying \( k \)-means algorithm to cluster a given set of data, in each iteration the \( k \)-means objective is guaranteed to decrease, except for when the algorithm has converged.

   \[ \text{True. See convergence property of \( k \)-means.} \]

c. (2 pts) If random variables \( A \) and \( B \) are not conditionally independent given \( C \), then they must not be independent from each other.

   \[ \text{False. See conditional independence notes.} \]

d. (2 pts) When applying SVM, removing examples that are not support vectors will not change the learned decision boundary.

   \[ \text{True. None-support vectors do not influence decision boundary.} \]

e. (2 pts) The two association rules \( A \rightarrow B \) and \( B \rightarrow A \) must have identical support.

   \[ \text{True. Supports of } A \rightarrow B \text{ and } B \rightarrow A \text{ are both support of } \{A \cup B\}. \]
3. (10 pts) Bayes classifier. For a two class problem with 5 binary features, how many parameters do we need to estimate for each of the following classifier:

- The basic Bayes Classifier that models the joint distribution of all five features.

\[
\begin{align*}
&P(Y=1) \quad P(Y=0) \quad 1 \text{ parameter (the other comes for free)}.
&P(CX|Y=1) \quad 2^5-1 \quad \text{parameters (number of possible X configurations-1)}.
&P(CX|Y=0) \quad 2^5-1 \quad \text{parameters}.
\end{align*}
\]

Total: \((2^5-1) \times 2 + 1 = 63\)

- The Naive Bayes Classifier that makes conditional independence assumption.

\[
\begin{align*}
&P(Y=1) \quad P(Y=0) \\
&P(CX_i|Y=1) \quad \text{for } i=1,\ldots,5. \quad 5 \text{ parameters}.
&P(CX_i|Y=0) \quad \text{for } i=1,\ldots,5 \quad 5 \text{ parameters}.
\end{align*}
\]

\(5 \times 2 + 1 = 11\)
4. (11 pts) Decision trees. Consider the following learning task with three features and a binary class label.

<table>
<thead>
<tr>
<th></th>
<th>V</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

a. (6pts) Build a decision stump using your favorite uncertainty measure (note that no calculator is needed if you use the training error here). What is the training error of this decision stump?

![Decision stump diagram]

b. (5pts) Suppose the true class concept is \( y = 1 \) if only if exactly one of the three features takes value 1. What would be the optimal decision tree for representing this concept?

![Optimal decision tree diagram]
5. (10 pts) SVM. Consider applying a soft margin SVM to the 1-dimensional dataset shown below. What will be the support vectors for each of the possible $c$ values?

\[ X_1 \quad X_2 \quad X_3 \quad X_4 \quad X_5 \quad X_6 \]

a. $c = 0$

The support vectors are: $X_1, X_2, X_3, X_4, X_5, X_6$.

The reason is:

No penalty for error. The tube will contain all examples inside making them all support vectors.

b. $c = \infty$

The support vectors are: $X_3, X_4$

The reason is:

Infinite penalty. Reduces to hard margin SVM.