Name: _________________________

- You have 110 minutes.
- There are 11 pages including cover page. Please make sure you are not missing any pages.
- Good luck!

<table>
<thead>
<tr>
<th></th>
<th>Max</th>
<th>Your pts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-midterm</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>Ensemble methods</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>GMM</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Learning theory</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>HAC</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>k-means</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>99</td>
<td></td>
</tr>
</tbody>
</table>
1. (35pts) **Pre-midterm**

   a. (4pts) For a neural network, which of the following choices most strongly affects the trade-off between under-fitting and over-fitting:
      
      i. The initial weights
      ii. The learning rate
      iii. The number of hidden nodes
      iv. The choice of the online or batch learning algorithm

   b. (6pts) When using generative models for classification, we assume the data is generated by a procedure defined by the generative model. Describe the generative procedure for text data when using a Naive Bayes classifier (using the bernoulli distribution).
c. (4pts) Consider soft margin SVM, described as the following optimization problem:

$$\arg \min_{W, b} |W|^2 + C \sum \xi_i$$

subject to \(y_i \cdot (W \cdot X_i + b) + \xi_i \geq 1\) and \(\xi_i \geq 0\) for \(i = 1, \ldots, n\)

This can be viewed as a penalty based method to avoid overfitting. Please specify which term in the objective function is the penalty term and which is the error term.

d. (7pts) Consider the following one dimensional data set. ‘x’ denotes negative examples and ‘o’ denotes positive examples. The exact data points and their labels are given in the table below. Please circle the support vectors, and mark the decision boundary \(w \cdot x + b = 0\) and solve for \(w\) and \(b\).

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>1</td>
<td>1.5</td>
<td>2.5</td>
<td>3</td>
<td>4</td>
<td>4.5</td>
</tr>
<tr>
<td>y</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ x \cdot (w \cdot x + b) = 0 \]
e. [6pts] Specify a neural network that represents boolean formula \((x_1 \land \neg x_2) \lor (\neg x_1 \land x_3)\).

f. (8pts) **True or False:**

i. Larger training sets lead to higher probability of overfitting.

ii. For the K-nearest neighbor algorithm, larger K values will lead to higher probability of overfitting.

iii. Logistic regression learns a non-linear decision boundary because the logistic function is non-linear.

iv. Given linearly separable data, online Perceptron only finds a local optimal (not globally optimal) solution of the hinge loss function, because different initializations of the weights can lead to different solutions.
2. (16pts) (**Ensemble methods**)

   a. (8pt) Consider a data set $D = x_1, x_2, ..., x_{10}$. We apply Adaboost with decision stump. Let $w_1, \cdots, w_{10}$ be the weights of the ten training examples respectively.

      i. (3pts) In the first iteration, $x_1, x_2$ and $x_3$ are misclassified. Please rank the updated weights $w_1, w_2, w_3, w_4$ in increasing order. Explain your ordering.

   ii. (5pts) In the second iteration, $x_3$ and $x_4$ are misclassified. Please rank the updated weights $w_1, w_2, w_3, w_4$ in increasing order. Explain your ordering.
b. (4pts) Which of the following algorithms do you expect to see most performance improvement when used with bagging? Briefly explain your answer in terms of stability of the algorithms.
   i. Decision stump
   ii. Decision tree with pruning
   iii. Decision tree without pruning

c. (4pts) If the training set contains noise in class labels, which ensemble learning method do you expect to be hurt more by the label noise, boosting or bagging? Why?

3. (8pts) (Gaussian Mixture Models) In the following figures, ‘+’ and ‘o’ denote positive and negative examples respectively. Three students fit Gaussian Mixture Models to this data.
   a. Students A and B fit one Gaussian to each class. The solid (dashed) ellipse indicates the probability contour for the positive (negative) class. Which model do you prefer and why? What constraints do they each use for the covariance matrix?
b. Student C fit two Gaussian to each class and used EM to learn the parameters. The left column below shows two initializations for EM and the right column shows the models after the 1st iteration. Match each initialization with its successive model.
4. (18pts) **Learning theory**

a. [5pts] Let $H_{2d}$ be the hypothesis space of all possible 2-dimensional linear threshold units. Prove that $\text{VC}(H_{2d}) \geq 3$.

b. [4pts] Suppose that you are able to find a set of 4 instances that cannot be shattered by $H_{2d}$. Does this imply that $\text{VC}(H_{2d}) < 4$? Explain.
c. [5pts] Is it always true that if a hypothesis space $H$ is parameterized by $d$ real numbers then $\text{VC}(H) \geq d$? If so prove your answer, if not give a counter example.

d. [4pts] In class we showed that if a consistent learning algorithm for a finite hypothesis space $H$ is provided with

$$m \geq \frac{1}{\epsilon} \left( \ln |H| + \ln \frac{1}{\delta} \right)$$

randomly drawn training instances, then we can state a certain guarantee. What is that guarantee? Make sure to clearly indicate the roles of $\epsilon$ and $\delta$. 
5. (10pts) (HAC)

a. [6pt] Draw the clustering dendrogram for the following sample of ten points in one dimension using single-link similarity.

   Sample = (−2.2, −2.0, −0.2, 0.1, 0.3, 0.4, 1.6, 1.7, 2.0, 2.1)

b. [5pt] Draw a two dimensional data set for which you would expect the results for single-link and complete-link to be substantially different. Explain.
6. (12pts) (k-Means Clustering and model selection)

a. [5pts] k-means is an algorithm that aims at finding $k$ clusters with minimum distortion. Write down the equation for the distortion of a given clustering. Make sure to properly define any notation you use.

b. [7pts] Consider the following strategy for selecting $k$, the number of clusters. Run k-means with different $k$ values, and choose $k$ that minimizes the distortion criterion. What is the potential problem of this strategy? How to use cross-validation, or the hold-out method for selecting $k$?