• You have 3 hours for the exam.

• The exam is closed book, closed notes except your one-page (two sides) or two-page (one side) crib sheet.

• Please use non-programmable calculators only.

• Mark your answers ON THE EXAM ITSELF. If you are not sure of your answer you may wish to provide a brief explanation. All short answer sections can be successfully answered in a few sentences AT MOST.

• For true/false questions, fill in the True/False bubble.

• For multiple-choice questions, fill in the bubbles for ALL CORRECT CHOICES (in some cases, there may be more than one). For a question with $p$ points and $k$ choices, every false positive will incur a penalty of $p/(k-1)$ points.

• For short answer questions, unnecessarily long explanations and extraneous data will be penalized. Please try to be terse and precise and do the side calculations on the scratch papers provided.

• Please draw a bounding box around your answer in the Short Answers section. A missed answer without a bounding box will not be regraded.

<table>
<thead>
<tr>
<th>First name</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Last name</td>
<td></td>
</tr>
<tr>
<td>SID</td>
<td></td>
</tr>
</tbody>
</table>

For staff use only:

| Q1. True/False | /23 |
| Q2. Multiple Choice Questions | /36 |
| Q3. Short Answers       | /26 |
| Total                   | /85 |
Q1. [23 pts] True/False

(a) [1 pt] Solving a non-linear separation problem with a hard margin Kernelized SVM (Gaussian RBF Kernel) might lead to overfitting.
   ○ True ○ False

(b) [1 pt] In SVMs, the sum of the Lagrange multipliers corresponding to the positive examples is equal to the sum of the Lagrange multipliers corresponding to the negative examples.
   ○ True ○ False

(c) [1 pt] SVMs directly give us the posterior probabilities \( P(y = 1|x) \) and \( P(y = -1|x) \).
   ○ True ○ False

(d) [1 pt] \( V(X) = E[X]^2 - E[X^2] \)
   ○ True ○ False

(e) [1 pt] In the discriminative approach to solving classification problems, we model the conditional probability of the labels given the observations.
   ○ True ○ False

(f) [1 pt] In a two class classification problem, a point on the Bayes optimal decision boundary \( x^* \) always satisfies \( P(y = 1|x^*) = P(y = 0|x^*) \).
   ○ True ○ False

(g) [1 pt] Any linear combination of the components of a multivariate Gaussian is a univariate Gaussian.
   ○ True ○ False

(h) [1 pt] For any two random variables \( X \sim \mathcal{N}(\mu_1, \sigma_1^2) \) and \( Y \sim \mathcal{N}(\mu_2, \sigma_2^2) \), \( X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \).
   ○ True ○ False

(i) [1 pt] Stanford and Berkeley students are trying to solve the same logistic regression problem for a dataset. The Stanford group claims that their initialization point will lead to a much better optimum than Berkeley’s initialization point. Stanford is correct.
   ○ True ○ False

(j) [1 pt] In logistic regression, we model the odds ratio (\( \frac{p}{1-p} \)) as a linear function.
   ○ True ○ False

(k) [1 pt] Random forests can be used to classify infinite dimensional data.
   ○ True ○ False

(l) [1 pt] In boosting we start with a Gaussian weight distribution over the training samples.
   ○ True ○ False

(m) [1 pt] In Adaboost, the error of each hypothesis is calculated by the ratio of misclassified examples to the total number of examples.
   ○ True ○ False

(n) [1 pt] When \( k = 1 \) and \( N \to \infty \), the kNN classification rate is bounded above by twice the Bayes error rate.
   ○ True ○ False

(o) [1 pt] A single layer neural network with a sigmoid activation for binary classification with the cross entropy loss is exactly equivalent to logistic regression.
   ○ True ○ False
(p) [1 pt] The loss function for LeNet5 (the convolutional neural network by LeCun et al.) is convex.
   ○ True  ○ False

(q) [1 pt] Convolution is a linear operation i.e. \((\alpha f_1 + \beta f_2) * g = \alpha f_1 * g + \beta f_2 * g\).
   ○ True  ○ False

(r) [1 pt] The k-means algorithm does coordinate descent on a non-convex objective function.
   ○ True  ○ False

(s) [1 pt] A 1-NN classifier has higher variance than a 3-NN classifier.
   ○ True  ○ False

(t) [1 pt] The single link agglomerative clustering algorithm groups two clusters on the basis of the maximum distance between points in the two clusters.
   ○ True  ○ False

(u) [1 pt] The largest eigenvector of the covariance matrix is the direction of minimum variance in the data.
   ○ True  ○ False

(v) [1 pt] The eigenvectors of \(AA^T\) and \(A^TA\) are the same.
   ○ True  ○ False

(w) [1 pt] The non-zero eigenvalues of \(AA^T\) and \(A^TA\) are the same.
   ○ True  ○ False
Q2. [36 pts] Multiple Choice Questions

(a) [4 pts] In linear regression, we model \( P(y|x) \sim \mathcal{N}(w^T x + w_0, \sigma^2) \). The irreducible error in this model is ________.

- \( \sigma^2 \)
- \( E[(y - E[y|x])^2|x] \)
- \( E[y|x] \)

(b) [4 pts] Let \( S_1 \) and \( S_2 \) be the set of support vectors and \( w_1 \) and \( w_2 \) be the learnt weight vectors for a linearly separable problem using hard and soft margin linear SVMs respectively. Which of the following are correct?

- \( S_1 \subseteq S_2 \)
- \( w_1 = w_2 \)
- \( S_1 \) may not be a subset of \( S_2 \)
- \( w_1 \) may not be equal to \( w_2 \).

(c) [4 pts] Ordinary least-squares regression is equivalent to assuming that each data point is generated according to a linear function of the input plus zero-mean, constant-variance Gaussian noise. In many systems, however, the noise variance is itself a positive linear function of the input (which is assumed to be non-negative, i.e., \( x \geq 0 \)). Which of the following families of probability models correctly describes this situation in the univariate case?

- \( P(y|x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y-(w_0+w_1x))^2}{2\sigma^2}\right) \)
- \( P(y|x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y-(w_0+w_1^2x))^2}{2\sigma^2}\right) \)
- \( P(y|x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y-(w_0+w_1\sigma^2x))^2}{2\sigma^2}\right) \)
- \( P(y|x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y-(w_0+w_1\sigma^2x))^2}{2\sigma^2}\right) \)

(d) [3 pts] The left singular vectors of a matrix \( A \) can be found in ________.

- Eigenvectors of \( AA^T \)
- Eigenvectors of \( A^T A \)
- Eigenvectors of \( A^2 \)
- Eigenvectors of \( AA^T \)

(e) [3 pts] Averaging the output of multiple decision trees helps ________.

- Increase bias
- Decrease bias
- Increase variance
- Decrease variance

(f) [4 pts] Let \( A \) be a symmetric matrix and \( S \) be the matrix containing its eigenvectors as column vectors, and \( D \) a diagonal matrix containing the corresponding eigenvalues on the diagonal. Which of the following are true:

- \( AS = SD \)
- \( AS = DS \)
- \( SA = DS \)
- \( AS = DS^T \)

(g) [4 pts] Consider the following dataset: \( A = (0, 2), B = (0, 1) \) and \( C = (1, 0) \). The k-means algorithm is initialized with centers at \( A \) and \( B \). Upon convergence, the two centers will be at

- \( A \) and \( C \)
- \( C \) and the midpoint of \( AB \)
- \( A \) and the midpoint of \( BC \)
- \( A \) and \( B \)
(h) [3 pts] Which of the following loss functions are convex?

- Misclassification loss
- Logistic loss
- Hinge loss
- Exponential Loss \( e^{(-yf(x))} \)

(i) [3 pts] Consider \( T_1 \), a decision stump (tree of depth 2) and \( T_2 \), a decision tree that is grown till a maximum depth of 4. Which of the following is/are correct?

- \( \text{Bias}(T_1) < \text{Bias}(T_2) \)
- \( \text{Bias}(T_1) > \text{Bias}(T_2) \)
- \( \text{Variance}(T_1) < \text{Variance}(T_2) \)
- \( \text{Variance}(T_1) > \text{Variance}(T_2) \)

(j) [4 pts] Consider the problem of building decision trees with \( k \)-ary splits (split one node into \( k \) nodes) and you are deciding \( k \) for each node by calculating the entropy impurity for different values of \( k \) and optimizing simultaneously over the splitting threshold(s) and \( k \). Which of the following is/are true?

- The algorithm will always choose \( k = 2 \)
- The algorithm will prefer high values of \( k \)
- There will be \( k - 1 \) thresholds for a \( k \)-ary split
- This model is strictly more powerful than a binary decision tree.
Q3. [26 pts] Short Answers

(a) [5 pts] Given that \((x_1, x_2)\) are jointly normally distributed with \(\mu = [\mu_1 \; \mu_2]\) and \(\Sigma = [\sigma_{11} \; \sigma_{12}; \sigma_{21} \; \sigma_{22}]\) \((\sigma_{21} = \sigma_{12})\), give an expression for the mean of the conditional distribution \(p(x_1|x_2 = a)\).

(b) [4 pts] The logistic function is given by \(\sigma(x) = \frac{1}{1+e^{-x}}\). Show that \(\sigma'(x) = \sigma(x)(1 - \sigma(x))\).

(c) Let \(X\) have a uniform distribution

\[
p(x; \theta) = \begin{cases} \frac{1}{\theta} & 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases}
\]

Suppose that \(n\) samples \(x_1, \ldots, x_n\) are drawn independently according to \(p(x; \theta)\).

(i) [5 pts] The maximum likelihood estimate of \(\theta\) is \(x(n) = \max(x_1, x_2, \ldots, x_n)\). Show that this estimate of \(\theta\) is biased.

(ii) [2 pts] Give an expression for an unbiased estimator of \(\theta\).
(d) [5 pts] Consider the problem of fitting the following function to a dataset of 100 points \( \{(x_i, y_i)\}, i = 1 \ldots 100: \)

\[
y = \alpha \cos(x) + \beta \sin(x) + \gamma
\]

This problem can be solved using the least squares method with a solution of the form:

\[
\begin{bmatrix}
\alpha \\
\beta \\
\gamma
\end{bmatrix} = (X^T X)^{-1} X^T Y
\]

What are \( X \) and \( Y \)?

\[ X = \quad Y = \]

(e) [5 pts] Consider the problem of binary classification using the Naive Bayes classifier. You are given two dimensional features \((X_1, X_2)\) and the categorical class conditional distributions in the tables below. The entries in the tables correspond to \( P(X_1 = x_1 | C_i) \) and \( P(X_2 = x_2 | C_i) \) respectively. The two classes are equally likely.

\[
\begin{array}{c|cc}
X_1 = & C_1 & C_2 \\
\hline
-1 & 0.2 & 0.3 \\
0 & 0.4 & 0.6 \\
1 & 0.4 & 0.1 \\
\end{array}
\quad
\begin{array}{c|cc}
X_2 = & C_1 & C_2 \\
\hline
-1 & 0.4 & 0.1 \\
0 & 0.5 & 0.3 \\
1 & 0.1 & 0.6 \\
\end{array}
\]

Given a data point \((-1, 1)\), calculate the following posterior probabilities:

\[
P(C_1 | X_1 = -1, X_2 = 1) =
\]

\[
P(C_2 | X_1 = -1, X_2 = 1) =
\]