• You have 1 hour 20 minutes for the exam.
• The exam is closed book, closed notes except your one-page 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.

| First name |          |          |
| Last name  |          |          |
| SID        |          |          |

For staff use only:

| Q1. True/False |          |
| Q2. Multiple Choice Questions | /21 |
| Q3. Short Answers | /15 |
| Total | /50 |
Q1. [14 pts] True/False

(a) [1 pt] In Support Vector Machines, we maximize $\frac{\|w\|^2}{2}$ subject to the margin constraints.
   ○ True    ● False

(b) [1 pt] In kernelized SVMs, the kernel matrix $K$ has to be positive definite.
   ○ True    ● False

(c) [1 pt] If two random variables are independent, then they have to be uncorrelated.
   ● True    ○ False

(d) [1 pt] Isocontours of Gaussian distributions have axes whose lengths are proportional to the eigenvalues of the covariance matrix.
   ○ True    ● False

(e) [1 pt] The RBF kernel ($K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2)$) corresponds to an infinite dimensional mapping of the feature vectors.
   ● True    ○ False

(f) [1 pt] If $(X, Y)$ are jointly Gaussian, then $X$ and $Y$ are also Gaussian distributed.
   ● True    ○ False

(g) [1 pt] A function $f(x, y, z)$ is convex if the Hessian of $f$ is positive semi-definite.
   ● True    ○ False

(h) [1 pt] In a least-squares linear regression problem, adding an $L_2$ regularization penalty cannot decrease the $L_2$ error of the solution $w$ on the training data.
   ● True    ○ False

(i) [1 pt] In linear SVMs, the optimal weight vector $w$ is a linear combination of training data points.
   ● True    ○ False

(j) [1 pt] In stochastic gradient descent, we take steps in the exact direction of the gradient vector.
   ○ True    ● False

(k) [1 pt] In a two class problem when the class conditionals $P(x|y = 0)$ and $P(x|y = 1)$ are modelled as Gaussians with different covariance matrices, the posterior probabilities turn out to be logistic functions.
   ○ True    ● False

(l) [1 pt] The perceptron training procedure is guaranteed to converge if the two classes are linearly separable.
   ● True    ○ False

(m) [1 pt] The maximum likelihood estimate for the variance of a univariate Gaussian is unbiased.
   ○ True    ● False

(n) [1 pt] In linear regression, using an $L_1$ regularization penalty term results in sparser solutions than using an $L_2$ regularization penalty term.
   ● True    ○ False
Q2. [21 pts] Multiple Choice Questions

(a) [2 pts] If $X \sim N(\mu, \sigma^2)$ and $Y = aX + b$, then the variance of $Y$ is:

- $a\sigma^2 + b$
- $a^2\sigma^2 + b$
- $a\sigma^2$
- $a^2\sigma^2$

(b) [2 pts] In soft margin SVMs, the slack variables $\xi_i$ defined in the constraints $y_i(w^T x_i + b) \geq 1 - \xi_i$ have to be:

- $< 0$
- $\leq 0$
- $> 0$
- $\geq 0$

(c) [4 pts] Which of the following transformations when applied on $X \sim N(\mu, \Sigma)$ transforms it into an axis aligned Gaussian? ($\Sigma = UDU^T$ is the spectral decomposition of $\Sigma$)

- $U^{-1}(X - \mu)$
- $(UD)^{-1}(X - \mu)$
- $UD(X - \mu)$
- $U(X - \mu)$
- $\Sigma^{-1}(X - \mu)$

(d) [2 pts] Consider the sigmoid function $f(x) = 1/(1 + e^{-x})$. The derivative $f'(x)$ is:

- $f(x) \ln f(x) + (1 - f(x)) \ln(1 - f(x))$
- $f(x)(1 - f(x))$
- $f(x)(1 + f(x))$

(e) [2 pts] In regression, using an $L_2$ regularizer is equivalent to using a ________ prior.

- Laplace, $2\beta \exp(-|x|/\beta)$
- Exponential, $\beta \exp(-x/\beta)$, for $x > 0$
- Gaussian with $\Sigma = cI, c \in R$
- Gaussian with diagonal covariance ($\Sigma \neq cI, c \in R$)

(f) [2 pts] Consider a two class classification problem with the loss matrix given as $\begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix}$. Note that $\lambda_{ij}$ is the loss for classifying an instance from class $j$ as class $i$. At the decision boundary, the ratio $\frac{P(y_2|x)}{P(y_1|x)}$ is equal to:

- $\frac{\lambda_{11} - \lambda_{22}}{\lambda_{21} - \lambda_{12}}$
- $\frac{\lambda_{11} - \lambda_{12}}{\lambda_{21} - \lambda_{12}}$
- $\frac{\lambda_{11} + \lambda_{22}}{\lambda_{21} + \lambda_{12}}$
- $\frac{\lambda_{11} - \lambda_{12}}{\lambda_{22} - \lambda_{21}}$

(g) [2 pts] Consider the $L_2$ regularized loss function for linear regression $L(w) = \frac{1}{2}\|Y - Xw\|^2 + \lambda\|w\|^2$, where $\lambda$ is the regularization parameter. The Hessian matrix $\nabla_w^2 L(w)$ is:

- $X^T X$
- $2\lambda X^T X$
- $X^T X + 2\lambda I$
- $(X^T X)^{-1}$

(h) [2 pts] The geometric margin in a hard margin Support Vector Machine is:

- $\frac{\|w\|^2}{2}$
- $\frac{1}{\|w\|^2}$
- $\frac{2}{\|w\|}$
- $\frac{2}{\|w\|^2}$

(i) [3 pts] Which of the following functions are convex?

- $\sin(x)$
- $|x|$
- $\min(f_1(x), f_2(x))$, where $f_1$ and $f_2$ are convex
- $\max(f_1(x), f_2(x))$, where $f_1$ and $f_2$ are convex
Q3. [15 pts] Short Answers

(a) [4 pts] For a hard margin SVM, give an expression to calculate \(b\) given the solutions for \(w\) and the Lagrange multipliers \(\{\alpha_i\}_{i=1}^N\).

Using the KKT conditions \(\alpha_i(y_i(w^T x_i + b) - 1) = 0\), we know that for support vectors, \(\alpha_i \geq 0\). Thus for some \(\alpha_i \geq 0\), \(y_i(w^T x_i + b) = 1\) and thus

\[
b = y_i - w^T x_i
\]

For numerical stability, we can take an average over all the support vectors.

\[
b = \sum_{x_i \in S_v} \frac{y_i - w^T x_i}{|S_v|}
\]

(b) Consider a Bernoulli random variable \(X\) with parameter \(p\) (\(P(X = 1) = p\)). We observe the following samples of \(X\): \((1, 1, 0, 1)\).

(i) [2 pts] Give an expression for the likelihood as a function of \(p\).

\[
L(p) = p^3(1-p)
\]

(ii) [2 pts] Give an expression for the derivative of the negative log likelihood.

\[
\frac{d\text{NLL}(p)}{dp} = \frac{1}{1-p} - \frac{3}{p}
\]

(iii) [1 pt] What is the maximum likelihood estimate of \(p\)?

\[
p = \frac{3}{4}
\]

c) [6 pts] Consider the weighted least squares problem in which you are given a dataset \(\{\tilde{x}_i, y_i, w_i\}_{i=1}^N\), where \(w_i\) is an importance weight attached to the \(i^{th}\) data point. The loss is defined as \(L(\beta) = \sum_{i=1}^N w_i(y_i - \beta^T x_i)^2\). Give an expression to calculate the coefficients \(\tilde{\beta}\) in closed form.

*Hint*: You might need to use a matrix \(W\) such that \(\text{diag}(W) = [w_1 \ldots w_N]^T\).

Define \(Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}\) and \(X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix}\).

Then \(L(\beta) = (Y - X\beta)^T W(Y - X\beta)\). Setting \(\frac{dL(\beta)}{d\beta} = 0\), we get

\[
\tilde{\beta} = (X^T WX)^{-1}X^T W Y
\]