

- You have 2 hours 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.
- 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). We have introduced a negative penalty for false positives for the multiple choice questions such that the expected value of randomly guessing is 0. Don't worry, for this section, your score will be the maximum of your score and 0, thus you cannot incur a negative score for this section.

| | |
|--|--|
| First name | |
| Last name | |
| SID | |
| First and last name of student to your left | |
| First and last name of student to your right | |

For staff use only:

| | |
|---|------|
| Q1. True or False | /10 |
| Q2. Multiple Choice | /24 |
| Q3. Decision Theory | /8 |
| Q4. Kernels | /14 |
| Q5. L2-Regularized Linear Regression with Newton's Method | /8 |
| Q6. Maximum Likelihood Estimation | /8 |
| Q7. Affine Transformations of Random Variables | /13 |
| Q8. Generative Models | /15 |
| Total | /100 |

Q1. [10 pts] True or False

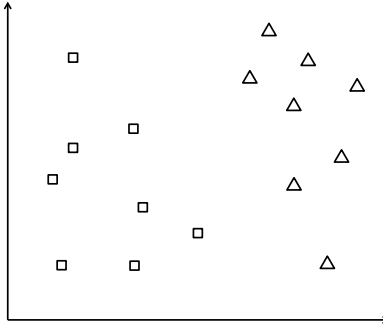
- (a) [1 pt] The hyperparameters in the regularized logistic regression model are η (learning rate) and λ (regularization term).
 True False
- (b) [1 pt] The objective function used in L2 regularized logistic regression is convex.
 True False
- (c) [1 pt] In SVMs, the values of α_i for non-support vectors are 0.
 True False
- (d) [1 pt] As the number of data points approaches ∞ , the error rate of a 1-NN classifier approaches 0.
 True False
- (e) [1 pt] Cross validation will guarantee that our model does not overfit.
 True False
- (f) [1 pt] As the number of dimensions increases, the percentage of the volume in the unit ball shell with thickness ϵ grows.
 True False
- (g) [1 pt] In logistic regression, the Hessian of the (non regularized) log likelihood is positive definite.
 True False
- (h) [1 pt] Given a binary classification scenario with Gaussian class conditionals and equal prior probabilities, the optimal decision boundary will be linear.
 True False
- (i) [1 pt] In the primal version of SVM, we are minimizing the Lagrangian with respect to w and in the dual version, we are minimizing the Lagrangian with respect to α .
 True False
- (j) [1 pt] For the dual version of soft margin SVM, the α_i 's for support vectors satisfy $\alpha_i > C$.
 True False

Q2. [24 pts] Multiple Choice

(a) [3 pts] Consider the binary classification problem where $y \in \{0, 1\}$ is the label and we have prior probability $P(y = 0) = \pi_0$. If we model $P(x|y = 1)$ to be the following distributions, which one(s) will cause the posterior $P(y = 1|x)$ to have a logistic function form?

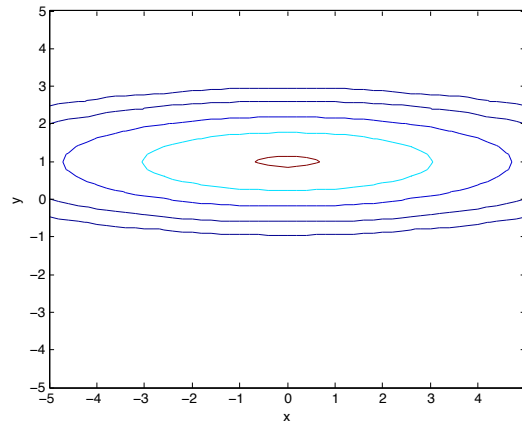
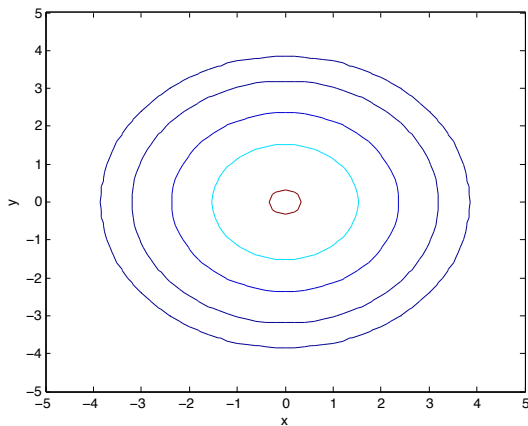
- Gaussian
 Uniform
 Poisson
 None of the above

(b) [3 pts] Given the following data samples (square and triangle belong to two different classes), which one(s) of the following algorithms can produce zero training error?



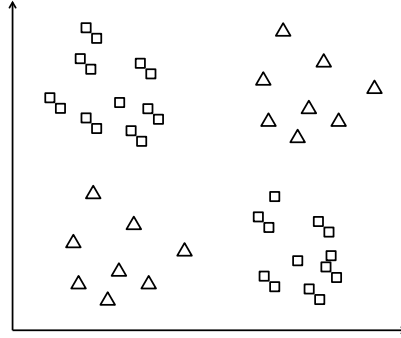
- 1-nearest neighbor
 Logistic regression
 Support vector machine
 Linear discriminant analysis

(c) [3 pts] The following diagrams show the iso-probability contours for two different 2D Gaussian distributions. On the left side, the data $\sim N(\mathbf{0}, \mathbf{I})$ where \mathbf{I} is the identity matrix. The right side has the same set of contour levels as left side. What is the mean and covariance matrix for the right side's multivariate Gaussian distribution?



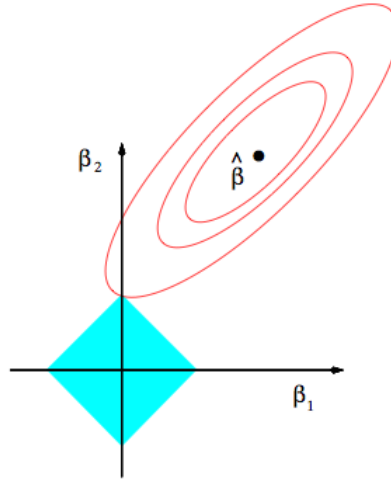
- $\mu = [0, 0]^T, \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
 $\mu = [0, 1]^T, \quad \Sigma = \begin{bmatrix} 4 & 0 \\ 0 & 0.25 \end{bmatrix}$
 $\mu = [0, 1]^T, \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
 $\mu = [0, 1]^T, \quad \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 0.5 \end{bmatrix}$

- (d) [3 pts] Given the following data samples (square and triangle mean two classes), which one(s) of the following kernels can we use in SVM to separate the two classes?



- Linear kernel
 Gaussian RBF (radial basis function) kernel
 Polynomial kernel
 None of the above

- (e) [3 pts] Consider the following plots of the contours of the unregularized error function along with the constraint region. What regularization term is used in this case?



- L_2
 L_∞
 L_1
 None of the above

- (f) [3 pts] Suppose we have a covariance matrix

$$\Sigma = \begin{bmatrix} 5 & a \\ a & 4 \end{bmatrix}$$

What is the set of values that a can take on such that Σ is a valid covariance matrix?

- $a \in \Re$
 $a \geq 0$
 $-\sqrt{20} \leq a \leq \sqrt{20}$
 $-\sqrt{20} < a < \sqrt{20}$

(g) [3 pts] The soft margin SVM formulation is as follows:

$$\begin{aligned} \min \quad & \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^N \xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i \quad \forall i \\ & \xi_i \geq 0 \quad \forall i \end{aligned}$$

What is the behavior of the width of the margin ($\frac{2}{\|\mathbf{w}\|}$) as $C \rightarrow 0$?

- Behaves like hard margin
- Goes to zero
- Goes to infinity
- None of the above

(h) [3 pts] In Homework 4, you fit a logistic regression model on spam and ham data for a Kaggle Competition. Assume you had a very good score on the public test set, but when the GSIs ran your model on a private test set, your score dropped a lot. This is likely because you overfitted by submitting multiple times and changing the following between submissions:

- λ , your penalty term
- ϵ , your convergence criterion
- η , your step size
- Fixing a random bug

(i) [0 pts] **BONUS QUESTION** (Answer this only if you have time and are confident of your other answers because this is not extra points.)

We have constructed the multiple choice problems such that every false positive will incur some negative penalty. For one of these multiple choice problems, given that there are p points, r correct answers, and k choices, what is the formula for the penalty such that the expected value of random guessing is equal to 0? (You may assume $k > r$)

Q3. [8 pts] Decision Theory

Consider the following generative model for a 2-class classification problem, in which the class conditionals are Bernoulli distributions:

$$\begin{aligned} p(\omega_1) &= \pi \\ p(\omega_2) &= 1 - \pi \\ x|\omega_1 &= \begin{cases} 1 & \text{with probability } 0.5 \\ 0 & \text{with probability } 0.5 \end{cases} \\ x|\omega_2 &= \begin{cases} 1 & \text{with probability } 0.5 \\ 0 & \text{with probability } 0.5 \end{cases} \end{aligned}$$

Assume the loss matrix

$$\begin{array}{cc} & \begin{array}{cc} \text{true class} = 1 & \text{true class} = 2 \end{array} \\ \begin{array}{c} \text{predicted class} = 1 \\ \text{predicted class} = 2 \end{array} & \left(\begin{array}{cc} 0 & \lambda_{12} \\ \lambda_{21} & 0 \end{array} \right) \end{array}$$

- (a) [8 pts] Give a condition in terms of λ_{12} , λ_{21} , and π that determines when class 1 should always be chosen as the minimum-risk class.

Q4. [14 pts] Kernels

- (a) [6 pts] Let k_1 and k_2 be (valid) kernels; that is, $k_1(\mathbf{x}, \mathbf{y}) = \Phi_1(\mathbf{x})^T \Phi_1(\mathbf{y})$ and $k_2(\mathbf{x}, \mathbf{y}) = \Phi_2(\mathbf{x})^T \Phi_2(\mathbf{y})$. Show that $k = k_1 + k_2$ is a valid kernel by explicitly constructing a corresponding feature mapping $\Phi(\mathbf{z})$.

- (b) [8 pts] The polynomial kernel is defined to be

$$k(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T \mathbf{y} + c)^d$$

where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, and $c \geq 0$. When we take $d = 2$, this kernel is called the quadratic kernel. Find the feature mapping $\Phi(\mathbf{z})$ that corresponds to the quadratic kernel.

Q5. [8 pts] L2-Regularized Linear Regression with Newton's Method

Recall that the objective function for L2-regularized linear regression is

$$J(\mathbf{w}) = \|X\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_2^2$$

where X is the design matrix (the rows of X are the data points).

The global minimizer of J is given by:

$$\mathbf{w}^* = (X^T X + \lambda I)^{-1} X^T \mathbf{y}$$

(a) [8 pts] Consider running Newton's method to minimize J .

Let \mathbf{w}_0 be an arbitrary initial guess for Newton's method. Show that \mathbf{w}_1 , the value of the weights after one Newton step, is equal to \mathbf{w}^* .

Q6. [8 pts] Maximum Likelihood Estimation

(a) [8 pts] Let x_1, x_2, \dots, x_n be independent samples from the following distribution:

$$P(x|\theta) = \theta x^{-\theta-1} \text{ where } \theta > 1, x \geq 1$$

Find the maximum likelihood estimator of θ .

Q7. [13 pts] Affine Transformations of Random Variables

Let \mathbf{X} be a d -dimensional random vector with mean $\boldsymbol{\mu}$ and covariance matrix Σ . Let $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$, where A is a $n \times d$ matrix and \mathbf{b} is a n -dimensional vector.

(a) [6 pts] Show that the mean of \mathbf{Y} is $A\boldsymbol{\mu} + \mathbf{b}$.

(b) [7 pts] Show that the covariance matrix of \mathbf{Y} is $A\Sigma A^T$.

Q8. [15 pts] Generative Models

Consider a generative classification model for K classes defined by the following:

- Prior class probabilities: $P(C_k) = \pi_k \quad k = 1, \dots, K$
- General class-conditional densities: $P(\mathbf{x}|C_k) \quad k = 1, \dots, K$

Suppose we are given training data $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$ drawn independently from this model.

The labels \mathbf{y}_i are “one-of- K ” vectors; that is, K -dimensional vectors of all 0’s except for a single 1 at the element corresponding to the class. For example, if $K = 4$ and the true label of \mathbf{x}_i is class 2, then

$$\mathbf{y}_i = [0 \quad 1 \quad 0 \quad 0]^T$$

(a) [5 pts] Write the log likelihood of the data set. You may use y_{ij} to denote the j^{th} element of \mathbf{y}_i .

(b) [10 pts] What are the maximum likelihood estimates of the prior probabilities?

(Hint: Remember to use Lagrange multipliers!)