CS 189 Spring 2015 Introduction to Machine Learning Final

- You have 2 hours 50 minutes for the exam.
- The exam is closed book, closed notes except your one-page (two-sided) cheat sheet.
- No calculators or electronic items.
- Mark your answers ON THE EXAM ITSELF. If you are not sure of your answer you may wish to provide a *brief* explanation and state your assumptions.
- For true/false questions, fill in the *True/False* bubble.
- For multiple-choice questions, fill in the bubble for **EXACTLY ONE** choice that represents the best answer to the question.

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	/44
Q2.	Multiple Choice	/33
Q3.	Decision Theory	/9
Q4.	Parameter Estimation	/8
Q5.	Locally Weighted Logistic Regression	/14
Q6.	Decision Trees	/7
Q7.	Convolutional Neural Nets	/11
Q8.	Streaming k-means	/9
Q9.	Low Dimensional Decompositions	/15
	Total	/150

Q1. [44 pts] True or False

(a) [2 pts] A neural network with multiple hidden layers and sigmoid nodes can form non-linear decision boundaries.

○ True ○ False

(b) [2 pts] All neural networks compute non-convex functions of their parameters.

○ True ○ False

(c) [2 pts] For logistic regression, with parameters optimized using a stochastic gradient method, setting parameters to 0 is an acceptable initialization.

○ True ○ False

(d) [2 pts] For arbitrary neural networks, with weights optimized using a stochastic gradient method, setting weights to 0 is an acceptable initialization.

○ True ○ False

(e) [2 pts] Given a design matrix $X \in \mathbb{R}^{n\times d}$, where $d \ll n$, if we project our data onto a k dimensional subspace using PCA where k equals the rank of X, we recreate a perfect representation of our data with no loss.

○ True ○ False

(f) [2 pts] Hierarchical clustering methods require a predefined number of clusters, much like k-means.

○ True ○ False

(g) [2 pts] Given a predefined number of clusters k, globally minimizing the k-means objective function is NP-hard.

○ True ○ False

(h) [2 pts] Using cross validation to select hyperparameters will guarantee that our model does not overfit.

○ True ○ False

(i) [2 pts] A random forest is an ensemble learning method that attempts to lower the bias error of decision trees.

○ True ○ False

(j) [2 pts] Bagging algorithms attach weights $w_1 ... w_n$ to a set of N weak learners. They re-weight the learners and convert them into strong ones. Boosting algorithms draw N sample distributions (usually with replacement) from an original data set for learners to train on.

○ True ○ False

(k) [2 pts] Given any matrix X, its singular values are the eigenvalues of $XX^\top$ and $X^\top X$.

○ True ○ False

(l) [2 pts] Given any matrix X, $(XX^\top + \lambda I)^{-1}$ for $\lambda \neq 0$ always exists.

○ True ○ False

(m) [2 pts] Backpropagation is motivated by utilizing Chain Rule and Dynamic Programming to conserve mathematical calculations.

○ True ○ False

(n) [2 pts] An infinite depth binary Decision Tree can always achieve 100% training accuracy, provided that no point is mislabeled in the training set.

○ True ○ False

(o) [2 pts] In One vs All Multi-Class Classification in SVM, we are trying to classify an input data point X as one of the N classes $(C_1 ... C_n)$, each of which has a parameter vector $\vec{w}_1 ... \vec{w}_n$. We classify point X as the class C_i which maximizes the inner product of X and $\vec{w}_i$. ○ True ○ False

(p) [2 pts] The number of parameters in a parametric model is fixed, while the number of parameters in a non-parametric model grows with the amount of training data.

○ True ○ False

(q) [2 pts] As model complexity increases, bias will decrease while variance will increase.

○ True ○ False

(r) [2 pts] Consider a cancer diagnosis classification problem where almost all of the people being diagnosed don't have cancer. The probability of correct classification is the most important metric to optimize.

○ True ○ False

(s) [2 pts] For the 1-Nearest Neighbors algorithm, as the number of data points increases to infinity in our dataset, the error of our algorithm is guaranteed to be bounded by twice the Bayes Risk.

○ True ○ False

(t) [2 pts] Increasing the dimensionality of our data always decreases our misclassification rate.

○ True ○ False

(u) [2 pts] It is possible to represent a XOR function with a neural network without a hidden layer.

○ True ○ False

(v) [2 pts] At high dimensionality, the KD tree speedup to the nearest neighbor can be slower than the naive nearest neighbor implementation.

○ True ○ False

Q2. [33 pts] Multiple Choice

(a) [3 pts] Given a Neural Net with N input nodes, no hidden layers, one output node, with Entropy Loss and Sigmoid Activation Functions, which of the following algorithms (with the proper hyper-parameters and initialization) can be used to find the global optimum?

○ Simulated Annealing (Gradient Descent with restarts)

○ Stochastic Gradient Descent

○ Mini-Batch Gradient Descent

○ Batch Gradient Descent

○ All of the above

○ None of the above

(b) [3 pts] Given function $f(x) = |x^2 + 3| - 1$ defined on $\mathbb{R}$:

○ Newtons Method on minimizing gradients will always converge to the global optimum in one iteration from any starting location

○ Stochastic Gradient Descent will always converge to the global optimum in one iteration

○ The problem is nonconvex, so it not feasible to find a solution.

○ All of the above

○ None of the above

(c) [3 pts] Daniel wants to minimize a convex loss function f(x) using stochastic gradient descent. Given a random starting point, mark the condition that would guarantee that stochastic gradient descent will converge to the global optimum. Let η_t = step size at iteration t.

○ $\eta_t < 0$

○ Constant step size η_t

○ Decreasing step size $\eta_t = \frac{1}{\sqrt{t}}$

○ Decreasing step size $\eta_t = \frac{1}{t^2}$

○ All of the above

○ None of the above

(d) [3 pts] Which of the following is true of logistic regression?

○ It can be motivated by ”log odds”

○ The optimal weight vector can be found using MLE.

○ It can be used with L1 regularization

○ All of the above

○ None of the above

(e) [3 pts] You've just finished training a decision tree for spam classification, and it is getting abnormally bad performance on both your training and test sets. You know that your implementation has no bugs, so what could be causing the problem?

○ Your decision trees are too shallow.

○ You need to increase the learning rate.

○ You are overfitting.

○ All of the above.

(f) [3 pts] The numerical output of a sigmoid node in a neural network:

○ Is unbounded, encompassing all real numbers.

○ Is unbounded, encompassing all integers.

○ Is bounded between 0 and 1.

○ Is bounded between -1 and 1.

(g) [3 pts] If n is the number of points in the training set, regular nearest neighbor (without KD trees, hashing, etc) has a classification runtime of:

○ $O(1)$

○ $O(\log n)$

○ $O(n)$

○ $O(n^2)$

(h) [3 pts] Consider the p-norm of a vector x defined using the notation $\|x\|_p$. Also note that α is a scalar. Which of the following is true?

○ $\|x\|_p + \|y\|_p \geq \|x + y\|_p$.

○ $\|\alpha x\|_p = |\alpha| \|x\|_p$.

○ $\|x\|_p = 0$ implies x is the zero vector.

○ All of the above.

(i) [3 pts] What are some practical problems with the sigmoidal activation function in neural nets?

○ It is convex, and convex functions cannot solve nonconvex problems

○ It does not work well with the entropy loss function

○ It can have negative values

○ Gradients are small for values away from 0, leading to the "Vanishing Gradient" problem for large or recurrent neural nets

(j) [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 which of the following between submissions: A) λ, your penalty term; B) η, your step size; C) ϵ, your convergence criterion; or D) Fixing a random bug:

○ A

○ B

○ A and B

○ A, B, and C

○ C and D

○ A, B, C, and D

(k) [3 pts] With access to an n-by-n matrix of pairwise data distances, but no access to the data itself, we can use which of the following clustering techniques: A) k-means; B) k-medoids; C) hierarchical clustering:

○ A

○ B

○ C

○ A and B

○ B and C

○ A, B, and C

(l) [0 pts] What was your favorite class of the ~~semester~~ ~~year~~ all time?

○ CS 189 - Introduction to Machine Learning

○ CS 189 - Classify EVERYTHING

○ CS 189 - Advanced MATLAB and Numpy

○ CS 189 - Kaggle Competitions for Dummies

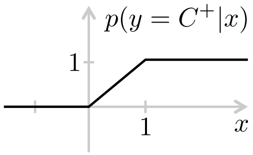
○ All of the above

○ ~~None of the above~~ (Choose this if you dare...)

Q3. [9 pts] Decision Theory

We are given a test that is designed to predict whether a patient y has cancer C^+ or not C^-. The test returns a value $x \in \mathcal{R}$ and we know the probability of the patient having cancer given the test results:

$$p(y = C^+|x) = \begin{cases} 0, & \text{if } x < 0 \\ x, & \text{if } 0 \leq x < 1 \\ 1, & \text{if } 1 \leq x \end{cases}$$



We also know that it is three times more costly to have a false negative than a false positive. Specifically, the loss matrix is:

Predicted:

Truth:		C^-	C^+
	C^-	0	10
	C^+	30	0

Suppose that we choose a fixed value x^*, and we predict C^+ if the test result is greater than x^* and C^- otherwise.

(a) [2 pts] What is the decision boundary (value of x) that minimizes the misclassification probability?

(b) [3 pts] What is the decision boundary (value of x^*) that minimizes the risk?

(c) [4 pts] If the test result is uniformly distributed in the interval $[-1, 1]$, what is the value of the minimum risk? Write your answer in terms of x^* (to avoid loss of points if your x^* is incorrect).

Q4. [8 pts] Parameter Estimation

Suppose you are given n observations, $X_1, ..., X_n$, independent and identically distributed with a Gamma(α, λ) distribution. The following information might be useful for one or more parts of the problem.

- If $X \sim$ Gamma(α, λ), then $\mathbb{E}[X] = \dfrac{\alpha}{\lambda}$ and $\mathbb{E}[X^2] = \dfrac{\alpha(\alpha+1)}{\lambda^2}$
- The probability density function of $X \sim$ Gamma(α, λ) is $f_X(x) = \dfrac{1}{\Gamma(\alpha)}\lambda^\alpha x^{\alpha-1}e^{-\lambda x}$ where the function Γ is only dependent on α and not λ.

The following notation might be useful for one or more parts of the problem: $\overline{X}_1 = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ and $\overline{X}_2 = \dfrac{1}{n}\sum_{i=1}^{n} X_i^2$.

(a) [4 pts] Find the estimators for α and λ using the method of moments. Remember you are trying to write α and λ as functions of the data.

(b) [4 pts] Suppose, we are given a known, fixed value for α. Compute the maximum likelihood estimator for λ.

Q5. [14 pts] Locally Weighted Logistic Regression

In this problem, we consider solving the problem of locally weighted logistic regression. Given data $\{(x_i, y_i) \in \mathbb{R}^d \times \{0,1\}\}_{i=1}^n$ and a query point x, we choose a parameter vector θ to minimize the loss (which is simply the negative log likelihood, weighted appropriately):

$$l(\theta; x) = -\sum_{i=1}^{n} w_i(x) \left[y_i \log(\mu(x_i)) + (1 - y_i) \log(1 - \mu(x_i)) \right]$$

where

$$\mu(x_i) = \frac{1}{1 + e^{-\theta \cdot x_i}}, \qquad w_i(x) = \exp\left(-\frac{\|x - x_i\|^2}{2\tau} \right)$$

where τ is a hyperparameter that must be tuned. Note that whenever we receive a new query point x, we must solve the entire problem again with these new weights $w_i(x)$.

Hint: the derivative of the logistic regression log likelihood with respect to θ is: $\sum_{i=1}^{n} (y_i - \mu(x_i)) x_i$

(a) [4 pts] Given a data point x, derive the gradient of $l(\theta; x)$ with respect to θ.

(b) [4 pts] Given a data point x, derive the Hessian of $l(\theta; x)$ with respect to θ.

(c) [2 pts] Given a data point x, write the update formula for gradient descent. Use the symbol η for an arbitrary step size.

(d) [2 pts] Given a data point x, write the update formula for Newton's method.

(e) [2 pts] Locally Weighted Logistic Regression is a

○ Parametric method ○ Nonparametric method

Q6. [7 pts] Decision Trees

Answer the following questions related to decision trees.

(a) [3 pts] In Homework 5, you first implemented a decision tree and then implemented a decision forest, which uses an ensemble method called bagging.

For neural network classification, it is typical to train k networks and average the results. Why not run your decision tree (using all of the data) k times and then average the results?

(b) [2 pts] True or false: Selecting the decision tree split (at each node as you move down the tree) that *minimizes classification error* will guarantee an optimal decision tree.

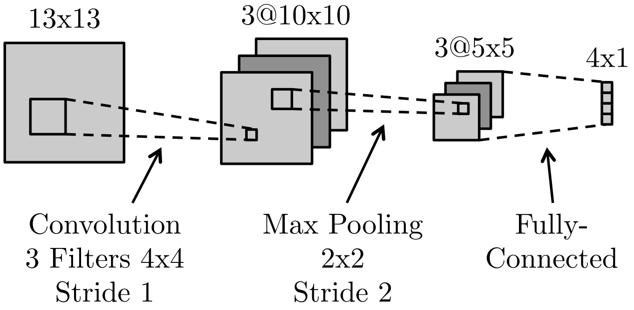
○ True ○ False

(c) [2 pts] True or false: Selecting the decision tree split (at each node as you move down the tree) that *maximizes information gain* will guarantee an optimal decision tree.

○ True ○ False

Q7. [11 pts] Convolutional Neural Nets

Below is a diagram of a small convolutional neural network that converts a 13x13 image into 4 output values. The network has the following layers/operations from input to output: convolution with 3 filters, max pooling, ReLu, and finally a fully-connected layer. For this network we will not be using any bias/offset parameters (b). Please answer the following questions about this network.



(a) [2 pts] How many weights in the convolutional layer do we need to learn?

(b) [2 pts] How many ReLu operations are performed on the forward pass?

(c) [2 pts] How many weights do we need to learn for the entire network?

(d) [2 pts] True or false: A fully-connected neural network with the same size layers as the above network (13x13 → 3x10x10 → 3x5x5 → 4x1) can represent any classifier that the above convolutional network can represent.

○ True ○ False

(e) [3 pts] What is the disadvantage of a fully-connected neural network compared to a convolutional neural network with the same size layers?

Q8. [9 pts] Streaming k-means

The standard k-means algorithm loads all data points altogether into the memory. In practice, data usually comes in a stream, such that they are sequentially processed and dropped (not stored in memory). The advantage of streaming algorithms is that their memory requirement is independent of the stream length. Thus, streaming algorithms are very useful in processing data that cannot fit into the memory.

In this problem, we will explore how to extend the k-means algorithm to process streaming data. Suppose that there are k clusters. The cluster centers are randomly initialized. Once the processor receives a data point $x \in \mathbb{R}^d$, it does the following:

1. Find the cluster whose center is the closest to x (in Euclidean distance), then add x to the cluster
2. Adjust the cluster center so that it equals the mean of all cluster members.

The algorithm outputs the k cluster centres after processing all data points in the stream.

According to the above algorithm specification, complete the streaming algorithm for k-means. Note that the algorithm's memory requirement should be independent of the stream length.

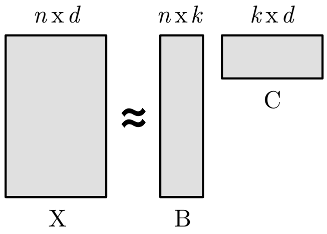
(a) [3 pts] List the variables that are stored in the memory and their initial values. Which variables should be the output of the algorithm?

(b) [3 pts] When the processor receives a data point x, state the updates that are made on the variables.

(c) [3 pts] In each iteration, suppose the processor receives a data point x along with its weight $w > 0$. We want the cluster center to be the weighted average of all cluster members. How do you modify the updates in question (b) to process weighted data?

Q9. [15 pts] Low Dimensional Decompositions

Given a design matrix $X \in \mathbb{R}^{n \times d}$ with $n > d$, we can create a low dimensional decomposition approximation $\tilde{X} = BC$, where $\tilde{X} \in \mathbb{R}^{n \times d}$, $B \in \mathbb{R}^{n \times k}$, $C \in \mathbb{R}^{k \times d}$, and $k < d$. The following figure shows a diagram of X approximated by B times C:



We can formulate several low dimensional techniques from CS 189 as solving the following optimization, subject to various constraints:

$$\min_{B,C} \|X - BC\|_F^2, \tag{1}$$

where $\| \cdot \|_F^2$ denotes the squared Frobenius norm of a matrix, that is, the sum of its squared entries.

(a) [2 pts] Which machine learning technique corresponds to solving (1) with constraint $\mathcal{C}_1$: each row of B is a vector e_i (a vector of all zeros, except a one in position i)?

○ k-means ○ k-medoids ○ SVD of X

(b) [3 pts] Describe the B and C matrices that result from solving (1) with constraint $\mathcal{C}_1$.

(c) [2 pts] Which machine learning technique corresponds to solving (1) with constraint $\mathcal{C}_2$: each column of B has norm equal to one?

○ k-means ○ k-medoids ○ SVD of X

(d) [3 pts] Describe the B and C matrices that result from solving (1) with constraint $\mathcal{C}_2$.

(e) [2 pts] Which machine learning technique corresponds to solving (1) with the constraints $\mathcal{C}_3$: each row of C is one of the rows from X and each row of B is a vector e_i (a vector of all zeros, except a one in position i)?

○ k-means ○ k-medoids ○ SVD of X

(f) [3 pts] Describe the B and C matrices that result from solving (1) with constraints $\mathcal{C}_3$.