ColumbiaX: CSMM.102x Machine Learning **Help**

Bookmarks

- Machine Learning Course: Getting Started
- Week 1
- Week 2
- Week 3
- Week 4
- Week 5
- Week 6
- Week 7
- Week 8
- Week 9
- Week 10
- Week 11
- Week 12
- Post-Course Survey
- **Final Exam**

Final Exam
Final Exam due Apr 18, 2017 05:00 IST

Final Exam > Final Exam > Final Exam

Final Exam

Bookmark this page

Question A

2.5/2.5 points (graded)

Check all instances of a supervised learning problem.

☑ separating spam from non-spam email using the text content of the email

☐ organizing people into groups based on a combination of their height, weight and age

☐ learning the topics from a corpus of documents

☑ predicting the presence/absence of a disease based on a blood test

✓

Submit You have used 1 of 1 attempt

✓ Correct (2.5/2.5 points)

Question A

2.5/2.5 points (graded)

If $x_1, \ldots, x_n$ are generated independent and identically distributed (i.i.d.) according to the distribution $p(x|\theta)$, then the joint likelihood can be written as

$$p(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta).$$

○ False

● True ✓

Submit You have used 1 of 1 attempt

✓ Correct (2.5/2.5 points)

Question A

3.5/3.5 points (graded)

You have data pairs $(y_i, x_i)_{i=1:n}$ where $x \in \mathbb{R}^{14}$ and you perform least squares linear regression to learn a function of the form $y = w_0 + x^T w$. What is the minimum number of samples required for this to be possible?

15 ✓

15

Submit You have used 1 of 1 attempt

✓ Correct (3.5/3.5 points)

Question A

4/4 points (graded)

Using the probabilistic approach to linear regression from Lecture 3, as well as the notations we have been using for the linear regression problem thus far, click all equivalent ways for generating from $p(y|X, w)$.

- ☑ $y_i = x_i^T w + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \quad \text{for} \;\; i = 1, \dots, n$
- ☑ $y_i \overset{ind}{\sim} N(x_i^T w, \sigma^2), \quad \text{for} \;\; i = 1, \dots, n$
- ☑ $y \sim N(Xw, \sigma^2 I)$
- ☑ $y = Xw + \vec{\epsilon}, \quad \vec{\epsilon} \sim N(0, \sigma^2 I)$

✓

Submit You have used 1 of 1 attempt

✓ Correct (4/4 points)

Question A

2.5/2.5 points (graded)

Given the model $y \sim N(Xw, \sigma^2 I)$, which of the following is true about the maximum likelihood estimator w_{ML}?

○ w_{ML} always has a unique solution

○ w_{ML} has the smallest variance among all estimators for w

○ $\mathbb{E}[w_{ML}] = (X^T X)^{-1} X^T y$

● w_{ML} is an unbiased estimator of w ✓

Submit You have used 1 of 1 attempt

✓ Correct (2.5/2.5 points)

Question A

2.5/2.5 points (graded)

Assume $w^* = \arg\min_w \|y - Xw\|_2^2 + \lambda g(w)$, which of the following is true?

○ When $g(w) = \|w\|^2$, the magnitude of values in w^* tend to increase

○ The solution for w^* is analytical for arbitrary positive function $g(w)$

● When $g(w) = \|w\|^2$, the values in w^* are more stable to variations in y and X ✓

○ The solution for w^* is always unique for arbitrary positive function $g(w)$

Submit You have used 1 of 1 attempt

✓ Correct (2.5/2.5 points)

Question A

2.5/2.5 points (graded)

The solution to ridge regression is $w_{RR} = (\lambda I + X^T X)^{-1} X^T y$. As λ increases, the value of $\|w_{RR}\|_2$

○ increases

● decreases ✓

Submit You have used 1 of 1 attempt

✓ Correct (2.5/2.5 points)

Question A

2.5/2.5 points (graded)

Which of the following are MAP solutions of a model with likelihood $p(y|w, X)$ and prior $p(w)$?

☑ $\arg\max_w \ln p(y, w|X)$

☑ $\arg\max_w \ln[p(y|w, X)p(w)]$

☐ $\arg\max_w \ln p(w|X)$

☐ $\arg\max_w \ln p(y|w, X)$

☑ $\arg\max_w \ln p(w|X, y)$

✓

Submit You have used 1 of 1 attempt

✓ Correct (2.5/2.5 points)

Question A

2.5/2.5 points (graded)

For a model with likelihood $p(y|w,X)$ and prior $p(w)$, given the training pairs (y,X) we test a new observation (y_0,x_0) by predicting y_0 given x_0. To compute this predictive distribution we need to calculate $p(y_0|w,x_0,y,X)$.

○ True

● False ✓

Submit You have used 1 of 1 attempt

✓ Correct (2.5/2.5 points)

Question A

0/2.5 points (graded)

For X an $n \times d$ matrix and y an n-dimensional vector, it is possible that the linear system $y = Xw$ may have multiple solutions when

☑ $n < d$

☐ $n \geq d$

☐ The null space of X is empty

☐ XX^T is invertible

✗

Submit You have used 1 of 1 attempt

✗ Incorrect (0/2.5 points)

Question A

4/4 points (graded)

Which of the following will likely give a sparse solution for w?

☑ $\arg\min_w \|y - Xw\|_2^2 + \lambda\|w\|_{1/2}$

☐ $\arg\min_w \|y - Xw\|_1 + \lambda\|w\|_2^2$

☐ $\arg\min_w \|y - Xw\|_2^2 + \lambda\|w\|_3^3$

☑ $\arg\min_w \|y - Xw\|_1 + \lambda\|w\|_{3/4}$

✓

Submit You have used 1 of 1 attempt

✓ Correct (4/4 points)

Question A

2.5/2.5 points (graded)
Which of the following describe a classification problem? (Check all that apply)

☐ predicting the gas milage of a car based on its weight and type

☑ predicting the presence of a disease based on preliminary tests

☑ predicting the monetary value of a check based on a photograph

☐ predicting the temperature tomorrow based on the temperature today

✓

Submit You have used 1 of 1 attempt

✓ Correct (2.5/2.5 points)

Question A

2.5/2.5 points (graded)
For a binary $\{-1, +1\}$ linear classifier, the coefficient vector w points in the direction of the _____ class.

○ -1

● $+1$ ✓

Submit You have used 1 of 1 attempt

✓ Correct (2.5/2.5 points)

Question A

2.5/2.5 points (graded)
Logistic regression is a model for regression.

○ TRUE

● FALSE ✓

Submit You have used 1 of 1 attempt

✓ Correct (2.5/2.5 points)

Question A

2.5/2.5 points (graded)
A discriminative classifier assumes a distribution on the covariates of a data set.

○ TRUE

● FALSE ✓

Submit You have used 1 of 1 attempt

✓ Correct (2.5/2.5 points)

Question A

2.5/2.5 points (graded)
The decision boundary for the logistic regression model is less sensitive to outliers than the least squares linear regression model.

● TRUE ✓

○ FALSE

Submit You have used 1 of 1 attempt

✓ Correct (2.5/2.5 points)

Question A

2.5/2.5 points (graded)
A feature expansion is useful when we want to

● learn a linear model in an alternate space ✓

○ learn a linear model in the original space

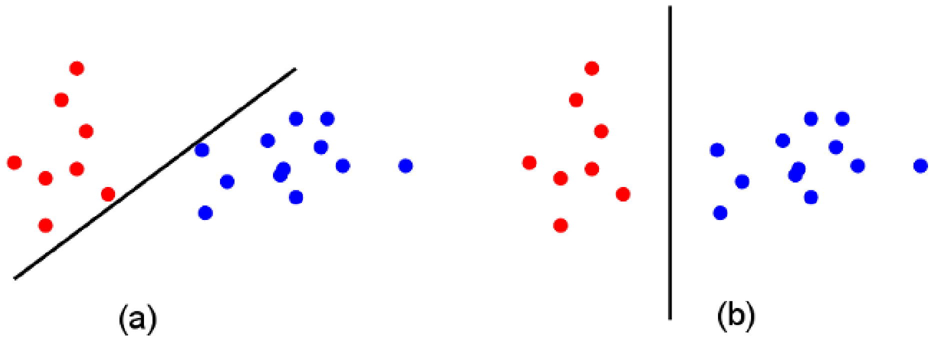
○ learn a linear model that evolves in time

Submit You have used 1 of 1 attempt

✓ Correct (2.5/2.5 points)

Question A

2.5/2.5 points (graded)



Which figure could correspond to the decision boundary of a support

Which figure could correspond to the decision boundary of a support vector machine?

○ (a)

● (b) ✓

○ both

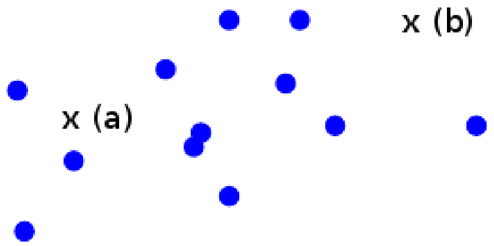
○ neither

Submit You have used 1 of 1 attempt

✓ Correct (2.5/2.5 points)

Question A

2.5/2.5 points (graded)



The figure contains a data set defined by the blue dots. Also shown in the figure are two locations marked by an "x" with a corresponding label. Which point(s), if any, are contained in the convex hull defined by the blue data points?

● (a) ✓

○ (b)

○ both

○ neither

Submit You have used 1 of 1 attempt

✓ Correct (2.5/2.5 points)

Question A

4/4 points (graded)

In a binary decision tree, every internal node has _____ children.

2 ✓

2

Submit You have used 1 of 1 attempt

✓ Correct (4/4 points)

Question A

2.5/2.5 points (graded)

When boosting a classifier, after round t the misclassified weights are multiplied by _______.

● e^{α_t} ✓

○ $e^{-\alpha_t}$

Submit You have used 1 of 1 attempt

✓ Correct (2.5/2.5 points)

Question A

4/4 points (graded)

For a new data point x_0, which of the following represents the boosted prediction of y_0?

● $y_0 = \text{sign}(\sum_t \alpha_t f_t(x_0))$ ✓

○ $y_0 = \text{sign}(\sum_t f_t(x_0))$

○ $y_0 = \text{sign}(\alpha f(x_0))$

○ $y_0 = \text{sign}(f(x_0))$

Submit You have used 1 of 1 attempt

✓ Correct (4/4 points)

Question A

2.5/2.5 points (graded)
True or False: The K-means objective function is convex and therefore the output of the K-means algorithm is the one true global solution.

○ TRUE

● FALSE ✓

Submit You have used 1 of 1 attempt

✓ Correct (2.5/2.5 points)

Question A

2.5/2.5 points (graded)
Check all probabilistic models.

☑ logistic regression

☐ support vector machines

☐ K-means

☑ Bayes classifiers

☐ Decision trees

✓

Submit You have used 1 of 1 attempt

✓ Correct (2.5/2.5 points)

Question A

2.5/2.5 points (graded)
The EM algorithm is guaranteed to monotonically increase the log likelihood.

● TRUE ✓

○ FALSE

Submit You have used 1 of 1 attempt

✓ Correct (2.5/2.5 points)

Question A

2.5/2.5 points (graded)
A mixture model represents the distribution of a data set as a weighted combination of simpler distributions.

● TRUE ✓

○ FALSE

Submit You have used 1 of 1 attempt

✓ Correct (2.5/2.5 points)

Question A

2.5/2.5 points (graded)

The maximum likelihood EM algorithm for the Gaussian mixture model will automatically learn an "appropriate" number of clusters for the data set by not assigning any data to the unnecessary clusters.

○ TRUE

● FALSE ✓

Submit You have used 1 of 1 attempt

✓ Correct (2.5/2.5 points)

Question A

2.5/2.5 points (graded)

Check all true statements about collaborative filtering using matrix factorization.

☑ we anticipate it will work because we make a low rank assumption

☐ all values in the matrix are needed before learning can begin

☑ probabilistic matrix factorization can be thought of as a set of connected ridge regression problems

☑ it can be thought of as a way for embedding users and objects into a latent space

✓

Submit You have used 1 of 1 attempt

✓ Correct (2.5/2.5 points)

Question A

2.5/2.5 points (graded)

True or false: Latent Dirichlet allocation can be thought of a a nonnegative matrix factorization problem matrix factorization problem.

● TRUE ✓

○ FALSE

Submit You have used 1 of 1 attempt

✓ Correct (2.5/2.5 points)

Question A

2.5/2.5 points (graded)
We made some loose connections between LDA and NMF. Which of the following is true?

○ LDA and NMF are both Bayesian models

○ LDA is a maximum likelihood model, while NMF can be thought of as a is a fully Bayesian model

● LDA is a fully Bayesian model, while NMF can be thought of as a maximum likelihood model ✓

○ neither LDA nor NMF have probabilistic interpretations

Submit You have used 1 of 1 attempt

✓ Correct (2.5/2.5 points)

Question A

2.5/2.5 points (graded)
True or False: The first principle component selects the direction of greatest variation in the data.

● TRUE ✓

○ FALSE

Submit You have used 1 of 1 attempt

✓ Correct (2.5/2.5 points)

Question A

4/4 points (graded)
A second-order Markov chain uses the previous _____ observations when making a prediction of the next observation.

2 ✓

Submit You have used 1 of 1 attempt

✓ Correct (4/4 points)

Question A

4/4 points (graded)
Enter the correct number: A 10x10 matrix can encode the transition probabilities of a ____-state Markov chain.

10 ✓

10

Submit You have used 1 of 1 attempt

✓ Correct (4/4 points)

Question A

2.5/2.5 points (graded)
In a hidden Markov model, the "hidden" portion corresponds to the ______.

○ observation sequence

● state transition sequence ✓

○ timestamp sequence

○ location sequence

Submit You have used 1 of 1 attempt

✓ Correct (2.5/2.5 points)

Question A

0/2.5 points (graded)
When we say "discrete HMM" the word "discrete" is referring to ______.

● a sequence indexed by a discrete set of time points. ✗

○ a sequence of discrete valued observations.

○ a sequence over a discrete set of hidden states.

Submit You have used 1 of 1 attempt

✗ Incorrect (0/2.5 points)

Question A

2.5/2.5 points (graded)
As discussed in class, in a continuous state Markov model, which of the following are not learned?

☑ the state transition distribution

☑ the observation distribution

☐ the hidden state sequence

☑ the initial state location

✓

Submit You have used 1 of 1 attempt

✓ Correct (2.5/2.5 points)