

CS 491: Introduction to Machine Learning
Spring 2015
Final Exam

Name: _____

User ID: _____

Instructions:

1. Write your name and user ID above. Do not begin the exam (look at other pages) until told to do so.
2. There should be 7 pages. Count the pages (without looking at the questions).
3. Do not discuss the exam with students who have not taken the exam!

Some useful formulas:

- $P(\mathbf{x}) = \sum_{\mathbf{y}} P(\mathbf{x}, \mathbf{y})$ (marginalization)
- $P(\mathbf{x}|\mathbf{y}) = \frac{P(\mathbf{x}, \mathbf{y})}{P(\mathbf{y})}$ (conditioning)
- $P(\mathbf{x}|\mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{x})P(\mathbf{x})}{P(\mathbf{y})}$ (Bayes theorem)
- $\mathbb{E}_{x \sim P}[f(X)] = \sum_x P(x)f(x)$ (Expectation)
- $X \sim \text{Normal}(\mu, \sigma) \implies P(X = x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- $X \sim \text{Multinoulli}(\theta) \implies P(\mathbf{x}) = \prod_{i=1}^K \theta_i^{x_i}$

	Points
Q1	/20
Q2	/20
Q3	/20
Q4	/20
Q5	/20
Total	

Q1. True-False questions (5 questions, 20 points total)

Circle correct answer. Give one sentence explanation or small picture.

Q1.1: (4 points) X independent of Y ($X \perp Y$) and Y independent of Z ($Y \perp Z$) implies that X is independent of Z ($X \perp Z$).

True / False. Explanation:

Q1.2: (4 points) Maximum a posteriori (MAP) estimation averages over the posterior parameter distribution to make predictions for new data.

True / False. Explanation:

Q1.3: (4 points) A logistic regression model with L_1 regularization will have lower training log loss / higher log likelihood than the same logistic regression model without regularization.

True / False. Explanation:

Q1.4: (4 points) The normalization term of a Markov random field, $\sum_{\mathbf{x}} e^{\psi(\mathbf{x})}$, with a tree graph structure connecting all n random variables can be computed in time linear in n .

True / False. Explanation:

Q1.5: (4 points) Consider a linear SVM with k support vectors. The SVM using polynomial kernel with degree $n > 2$ trained from the same data must have at least k support vectors.

True / False. Explanation:

Q2. Short Answer (3 questions, 20 points total)

Q2.1: (6 points) Which model is more prone to overfitting: a decision tree of depth k or a logistic regression model with k features? Why?

Q2.2: (8 points) Consider a binary logistic regression model, $P(Y = 1|x) = \frac{2^{w_2 x_2 + w_1 x_1 + w_0}}{2^{w_2 x_2 + w_1 x_1 + w_0} + 1}$, that provides predictions of: $P(Y = 1|X_1 = 2, X_2 = 0) = 0.5$, $P(Y = 1|X_1 = 0, X_2 = 3) = 0.5$, and $P(Y = 1|X_1 = 0, X_2 = 0) = \frac{2}{3}$. What are the values of w_0 , w_1 , and w_2 ?

$$w_0 =$$

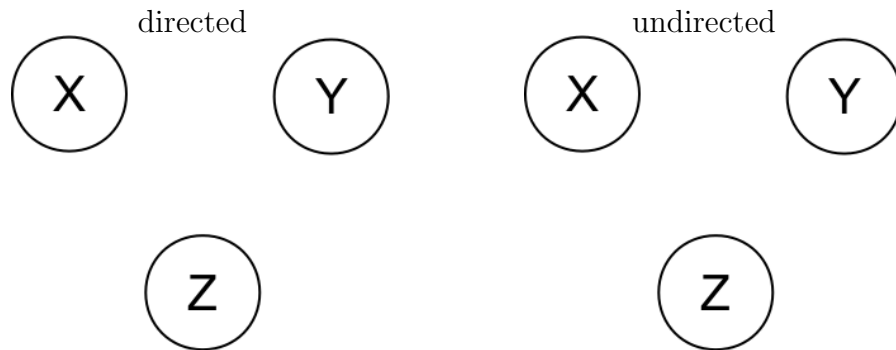
$$w_1 =$$

$$w_2 =$$

Q2.3: (6 points) Consider the expectation-maximization algorithm for Naïve Bayes with m unlabeled examples and a single input x . Let $\alpha_i(y)$ represent the expectation step's computation of $P(Y = y|x)$. What is the M step's conditional probability estimate for $\hat{P}(x|y)$ in terms of $\alpha_i(y)$ and training data? (Hint: for labeled data, $\hat{P}(x|y) = \frac{\sum_{i=1}^m I(X_i = x_i \wedge Y_i = y_i)}{\sum_{i=1}^m I(Y_i = y_i)}$, where $I(\cdot)$ is an indicator function returning 1 when the function is true and 0 otherwise.)

Q3. Graphical Models (20 points total)

Q3.1: (5 points) Given three random variables, X , Y , and Z , draw a **directed graphical model (Bayesian network)** (left) and an **undirected graphical model (Markov random field)** (right) that share at least one (conditional) independence property and each have at least one unique (conditional) independence property.



Q3.2: (4 points) Write a (conditional) independence property that both graphical models share (e.g., $X \perp Z | Y$):

Q3.3: (4 points) Write a (conditional) independence property that the directed graphical model possesses that the undirected graphical model does not:

Q3.4: (4 points) Write a (conditional) independence property that the undirected graphical model possesses that the directed graphical model does not:

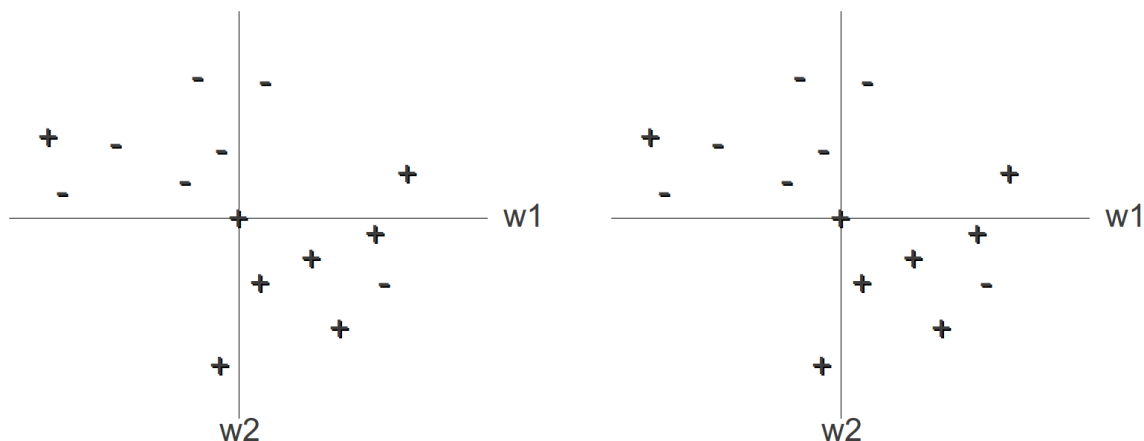
Q3.5: (4 points) Assuming that X , Y , and Z are binary-valued. Write a joint probability distribution that corresponds to your undirected graphical model above and has no additional independence properties.

Q4. Support Vector Machines (20 points)

Q4.1 Linear SVM Consider the linear support vector machine,

$$\min_{\mathbf{w}, b, \xi \geq 0} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \xi_i \text{ such that: } y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1 - \xi_i \quad \forall i \in \{1, \dots, n\},$$

on the following dataset:



- (5 points)** On the left, plot the decision boundary and margin boundaries when C is very large (as C goes to infinity) and circle the examples serving as support vectors.
- (5 points)** On the right, plot the decision boundary and margin boundaries when C is more moderate and circle the examples serving as support vectors.

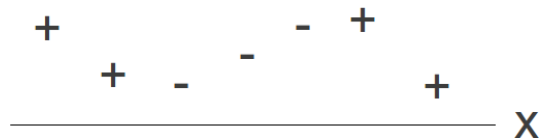
Q4.2 Kernelization Consider using a feature function ϕ that transforms from the original input space (size $|X|$) to a feature space of size $|F|$ with a corresponding kernel function K that requires k time to evaluate. Further, assume that the size of the training data is m and the number of support vectors is $|S|$.

- (5 points)** What is the computational complexity of making a prediction for a new example \mathbf{x} using the linear support vector machine without the kernel trick?
- (5 points)** What is the computational complexity of making a prediction for a new example \mathbf{x} using the linear support vector machine with the kernel trick?

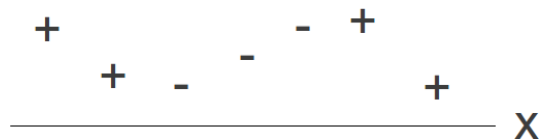
Q5. Boosting (20 points)

Consider the AdaBoost algorithm with threshold functions ($x_i > c \implies x_i = '+'$ and '-' otherwise or with labels reversed) as weak classifiers.

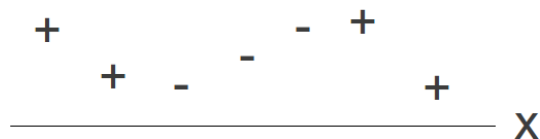
Q5.1: (5 points) Draw the first weak classifier learned on the following dataset. Circle the incorrectly labeled examples that will be more heavily weighted.



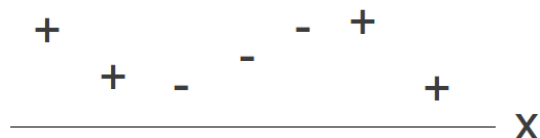
Q5.2: (5 points) Draw the second weak classifier learned on the resulting weighted dataset. Circle the incorrectly labeled examples that will be more heavily weighted.



Q5.3: (5 points) Draw the third weak classifier learned on the resulting weighted dataset. Circle the incorrectly labeled examples that will be more heavily weighted.



Q5.3: (5 points) Draw the decision functions and combined weights for each region that would produce a minimum amount of training set errors. (Which will be larger, α_1 or α_2 ?)



Extra space