CS 491: Introduction to Machine Learning
Spring 2013
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 10 pages. Count the pages (without looking at the questions).
3. Q1 contains multiple choice problems. Circle every answer that you believe is correct.
4. 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	/15
Q3	/20
Q4	/20
Q5	/15
Q6	/10
Total	

Q1. True-False questions (10 questions, 20 points total)

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

Q1.1: (2 points) For a continuous random variable x and its probability distribution $P(x)$, it hold that $0 \leq P(x) \leq 1$. True / False.

Q1.2: (2 points) Given a reasonable prior distribution, Bayesian parameter estimation and frequentist estimation (maximum likelihood) converge to the same solution given infinite data. True / False.

Q1.3: (2 points) Consider a decision tree with input features $X_1, X_2, ..., X_n$ and class label Y. If X_2 is independent of Y ($X_2 \perp Y$), then no decision based on X_2 will appear in the decision tree. True / False.

Q1.4: (2 points) L_1 regularization, $\lambda \sum_i |w_i|$ provides sparser learned models (i.e., fewer non-zero parameters) than L_2 regularization, $\lambda \sum_i w_i^2$. True / False.

Q1.5: (2 points) The hinge loss (from Support Vector Machines) is a lower bound on the 0-1 loss. True / False.

Q1.6: (2 points) Decision boundaries for the SVM using the "kernel trick" are always linear in the original feature space. True / False.

Q1.7: (2 points) The back-propagation algorithm learns a globally optimal neural network with hidden layers. True / False.

Q1.8: (2 points) In importance sampling, the proposal distribution, $q(x)$, scaled by M must upper bound the target distribution $p(x)$. True / False.

Q1.9: (2 points) Markov chain Monte Carlo requires conjugate prior distributions to generate samples from a probability distribution. True / False.

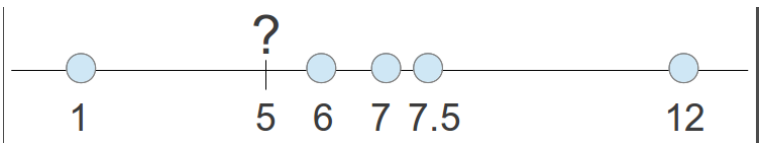
Q1.10: (2 points) The weak classifier weights α_i in AdaBoost are always non-negative. True / False.

Q2. Short Answer (3 questions, 15 points total)

Q2.1: (5 points) What are some reasons for preferring discriminative prediction techniques (e.g., SVM or logistic regression) over generative prediction techniques (e.g., naive Bayes)? What are reasons for preferring generative prediction techniques over discriminative techniques? (For what tasks would one or the other be needed?)

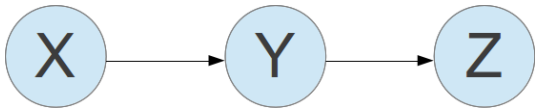
Q2.2: (5 points) What are the purposes of the training set, validation set, and testing set in the construction and evaluation of a machine learning technique, such as the neural network? When is each used and for what purpose?

Q2.3: (5 points) Consider kernel density estimation for the following five datapoints using the box kernel with bandwidth w. For what choice of kernel bandwidth is the density of the point at 5 maximized? Draw the resulting density estimate for the entire domain.

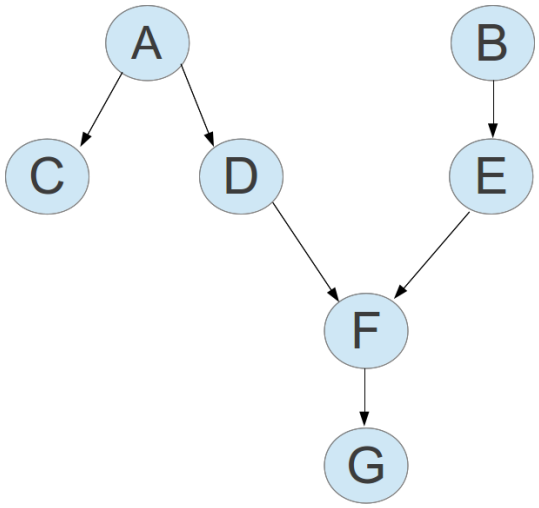


Q3. Bayesian Networks (20 points total)

Q3.1: (5 points) Write the expression for $P(X = x|Z = z)$ in terms of the conditional probability terms of the Bayesian network.



Q3.2: (10 points) Consider the Bayesian network for the following set of questions:



True or False, $A \perp B$?

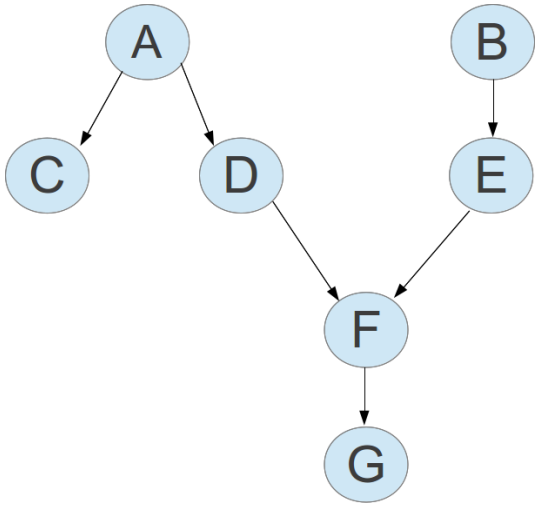
True or False, $A \perp E|D$?

True or False, $B \perp C|G$?

True or False, $B \perp D$?

True or False, $C \perp G$?

Q3.3: (5 points) Again, consider the Bayesian network:



Assume the variables can take $|A|, |B|, |C|, |D|, |E|, |F|, |G|$ values.
How many additions and multiplications are needed to efficiently compute $P(C = c)$? **(2 points)**

How many additions and multiplications are needed to efficiently compute $P(G = g)$? **(3 points)**

Q4. Undirected Graphical Models (20 points total)

Consider the Markov random field defined over binary variables X_1, X_2, X_3, with distribution

$$P(x_1, x_2, x_3) \propto e^{\theta_1 x_1 x_2 + \theta_2 x_1 x_3 + \theta_3 x_2 x_3} \tag{1}$$

Q4.1: **(5 points)** Given that $\theta_1 = \theta_2 = \theta_3 = 1$, what is the probability of $P(X_1 = 1, X_2 = 1, X_3 = 1)$?

Q4.2: **(10 points)** Given two training examples, $(1, 1, 0)$ and $(0, 0, 1)$, what choice of θ_1, θ_2, θ_3 parameters maximize the data likelihood? (Hint: two parameters should be easy to determine; the third should require calculus)

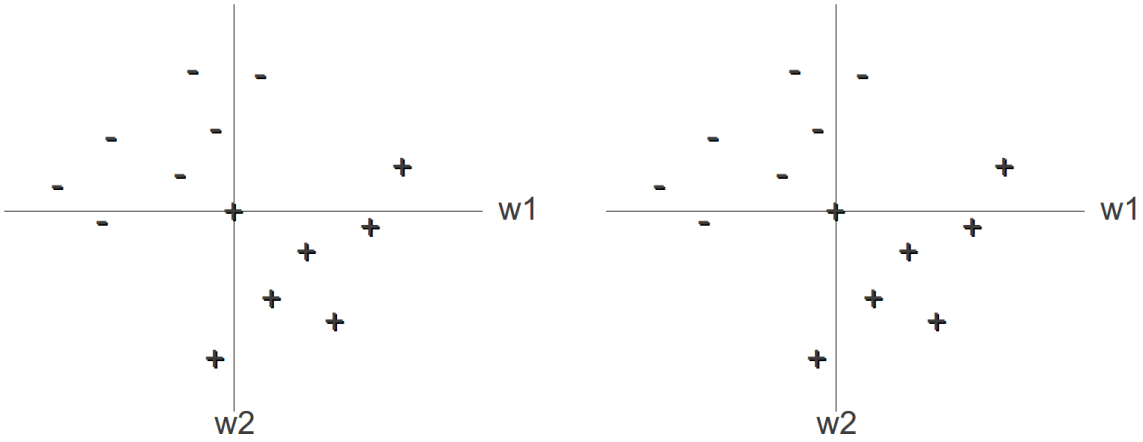
Q4.3: **(5 points)** In addition to the existing features, x_1x_2, x_1x_3, x_2x_3, including which pairwise boolean function as a feature $f(x_i, x_j)$ would most greatly improve the data likelihood of the training examples in Q4.2?

Q5. Support Vector Machines (and Logistic Regression) (15 points)

Q5.1 (6 points) Consider the linear support vector machine,

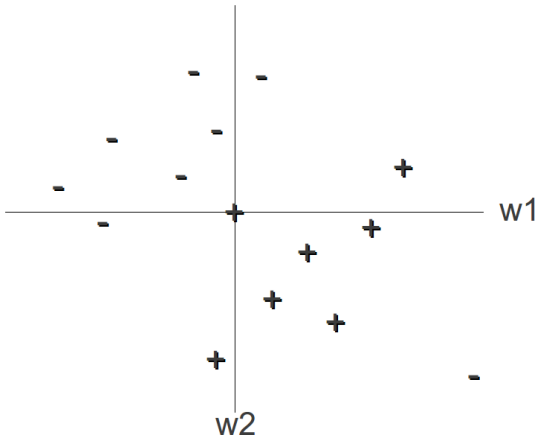
$$\min_{\mathbf{w},\epsilon,\xi} ||\mathbf{w}||^2 + C\sum_{i=1}^{n} \xi_i$$
$$\text{such that: } y_i(\mathbf{w}\cdot\mathbf{x}_i - b) \geq 1 - \xi_i \qquad \forall i \in \{1, ..., n\},$$

on the following dataset:

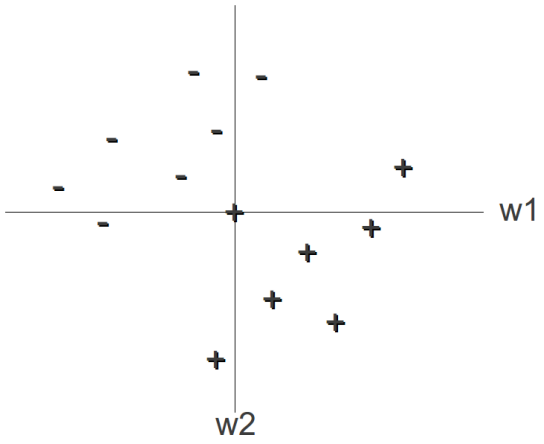


On the left, plot the decision boundary and margins when C is very large (as C goes to infinity). On the right, plot the decision boundary and margins when C is more moderate.

Q5.2 (4 points) On the following data set with an anomolous '-' example, is the logistic regression model or the support vector machine (with moderate C) more sensitive to this anomolous example? Plot and label the logistic regression and SVM decision boundaries.



Q5.3 (4 points) Consider the reduced logistic regression model, $P(Y|\mathbf{X}) \propto e^{w_1 x_1 + w_2 x_2}$ (without constant offset). What affect will changing the label of the training example at the origin $(0, 0)$ from '+' to '-' have on the learned logistic regression model? Plot the "before" and "after" decision boundaries.



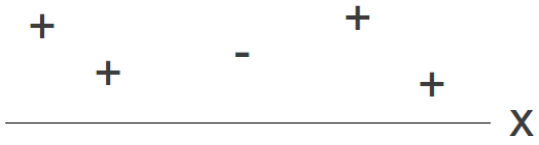
Q6. Boosting (10 points)

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

Q6.1: (3 points) Draw the first weak classifier learned on the following dataset. Circle the incorrectly labeled examples that will be more heavily weighted.

Q6.2: (3 points) Draw the second weak classifier learned on the resulting weighted dataset. Circle the incorrectly labeled examples that will be more heavily weighted.

Q6.3: (4 points) In how many iterations does the algorithm perfectly label the training data (if at all)? Draw the resulting decision function and combined weights for each region.



Extra space