# CS 491: Introduction to Machine Learning
## Spring 2014
## 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. There is no penalty for guessing.
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 | /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)** 3-Nearest Neighbor is guaranteed to have a lower training set error (not cross-validated) than 5-Nearest Neighbor.

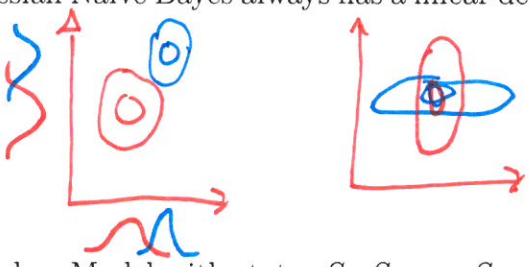
True / False

Explanation:

**Q1.2: (4 points)** Multi-variate Gaussian Naïve Bayes always has a linear decision boundary.

True / False

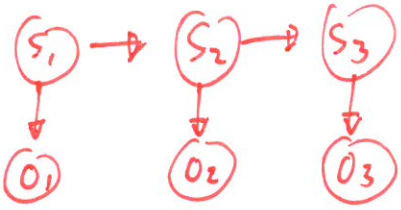
Explanation:



**Q1.3: (4 points)** In the Hidden Markov Model with states $S_1, S_2, \ldots, S_T$ and observations $O_1, O_2, \ldots, O_T$, the following independence property holds: $O_t \perp O_{t+1} | S_t$.

True / False

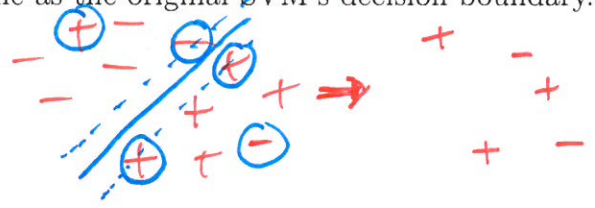
Explanation:




**Q1.4: (4 points)** If the non-support examples of a trained SVM are removed from the training set and the SVM is retrained from the smaller data, the decision boundary is guaranteed to be the same as the original SVM's decision boundary.

True / False

Explanation:



**Q1.5: (4 points)** The first $N$ samples (for some reasonably large $N$ based on the distribution) in the chain produced by the Metropolis-Hastings algorithm should be discarded.

True / False

Explanation: Burn in to reach a ~~world~~ probable region of the distribution

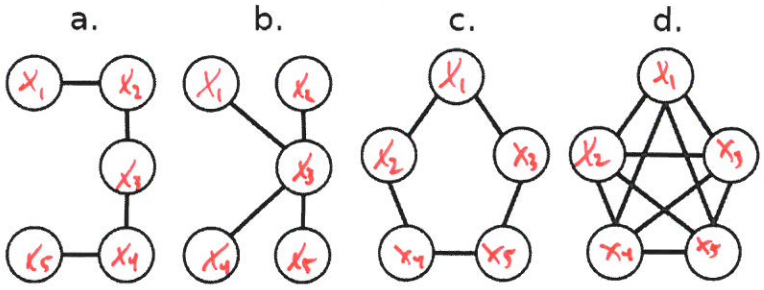
## Q2. Short Answer (4 questions, 20 points total)

**Q2.1: (4 points)** Given a classification method with some model parameters that we need to choose (e.g., the choice of $k$ for $k$-nearest neighbors, regularization parameters in logistic regression), how should the performance on unseen data be estimated using an available data set? Describe specifically how the available data should be used.

withhold test data

train data ⇒ k-fold cross-validation

**Q2.2: (8 points)** Consider the following four graphical models (Markov random fields).




What is the time complexity of variable elimination on each of these graphs in terms of the number of variables, $n$, and the number of values each can take, $|X|$?

Choose from: $O(n|X|)$, $O(n^2|X|)$, $O(n|X|^2)$, $O(n^3|X|)$, $O(n^2|X|^2)$, $O(n|X|^3)$, $O(n^{|X|})$, and $O(|X|^n)$ time.

$\psi_1(X_2) = \sum_{X_1} e^{\quad}$ $O(|X|^2)$

a. $\Rightarrow$ $O(n|X|^2)$

b. $\sum_{X_3} \psi_1(X_3)\,\psi_2(X_3)\,\psi_4(X_3)\,\psi_5(X_3)$ $\rightarrow O(n|X|^2)$

$O((n-1)|X|^2 + |X| \cdot (n-1))$

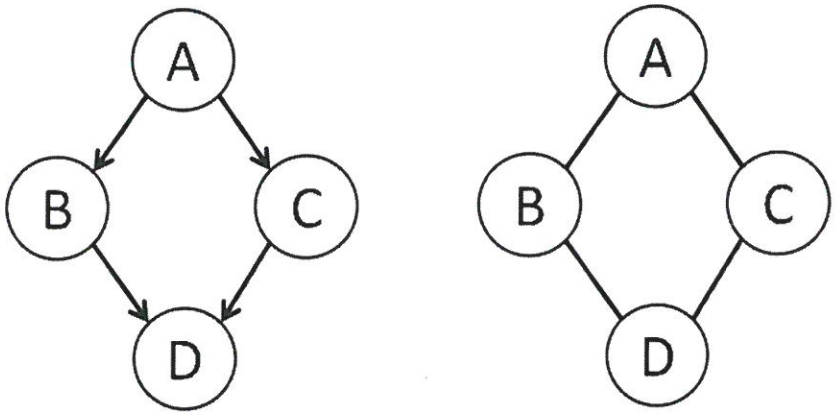
c.

d. $\psi_1(X_2, X_3) = \sum_{X_1} e^{f(X_1, X_2, X_3)}$ $O(|X|^3) \Rightarrow O(n|X|^3)$

$\psi_1(X_2, X_3, X_4, X_5) = \sum_{X_1} e^{f(X_1, X_2, X_3, X_4, X_5)}$ $O(|X|^n)$

# Q3. Graphical Models (20 points total)

Consider the following Bayesian Network and Markov Random Field:




**Q3.1**: **(5 points)** What is one independence property (e.g., $X \perp Y|Z$) that the Bayesian Network possesses that the Markov Random Field does not? (~~Hint: (conditional) probabilities must sum to one, providing on fewer free parameter.~~)

$B \perp C \mid A$

**Q3.2**: **(5 points)** What is one independence property (e.g., $X \perp Y|Z$) that the Markov Random Field possesses that the Bayesian Network does not? (~~Hint: probabilities still sum to one.~~)

$B \perp C \mid A, D$

**Q3.3**: **(5 points)** Assuming that each random variables has $k$ values, what is the maximum number of free parameters that the Bayesian Network can possess?

$P(A) \rightarrow (k-1)$ $\quad P(D|B,C) \rightarrow k^2(k-1)$

$P(B|A) \rightarrow k(k-1)$

$p(C|A) \rightarrow k(k-1)$

$\Rightarrow k^2(k-1) + 2k(k-1) + (k-1)$

**Q3.4**: **(5 points)** Assuming that each random variables has $k$ values, what is the maximum number of free parameters that Markov Random Field can possess?

$\Psi(a,b) \rightarrow k^2$ $\quad$ but distribution must sum to 1

$\psi(a,c) \rightarrow k^2$

$\psi(b,d) \rightarrow k^2$

$\psi(c,d) \rightarrow k^2$

$\Rightarrow 4k^2 - 1$

# Q4. Logistic Regression (20 points)

Consider the logistic regression model with parameters $w_0, w_1, w_2$:

$$P(y=1|x_1,x_2) = \frac{2^{w_2x_2+w_1x_1+w_0}}{1+2^{w_2x_2+w_1x_1+w_0}}. \quad = .5 \qquad (1)$$

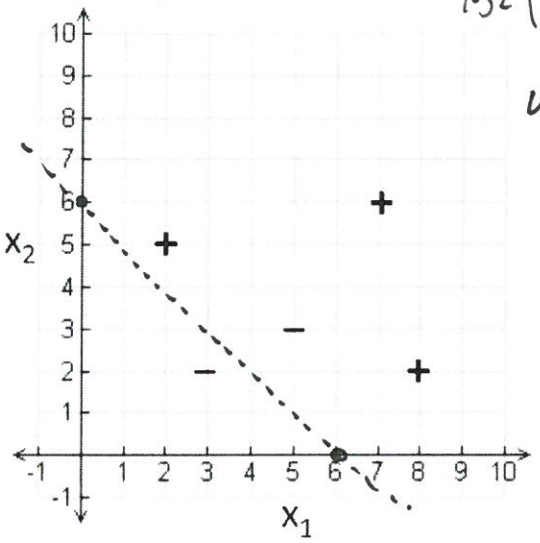
Note, base 2 is employed rather than $e$ for computational convenience.

**Q4.1 (5 points)** For parameter weights $w_0 = -6$, $w_1 = 1$, $w_2 = 1$, draw the decision boundary on the following plot:

$\log_2\left(\frac{1}{2}\cdot 2^{w_2x_2+w_1x_1+w_0}\right) = \log_2(1)$

$w_2x_2+w_1x_1+w_0 = 0$

$x_2+x_1-6=0$




**Q4.2 (5 points)** What is the log likelihood of the negative datapoint at $(x_1 = 5, x_2 = 3)$, $\log_2 P(Y=0|x_1=5,x_2=3)$ in the logistic regression model from Q4.1?

$$\log_2 P(y=0|x_1,x_2) = \log_2 \frac{1}{1+2^{w_2x_2+w_1x_1+w_0}} = -\log\left(1+2^{w_2x_2+w_1x_1+w_0}\right)$$

$1\cdot 3+1\cdot 5-6=2$

**Q4.3 (5 points)** What is the gradient of this datapoint?

$-\log_2(5)$

$\frac{\partial}{\partial w_0}\log_2 P(Y=0|x_1=5,x_2=3) = -4/5$

$\frac{\partial}{\partial w_1}\log_2 P(Y=0|x_1=5,x_2=3) = -4/5\cdot 5$

$\frac{\partial}{\partial w_2}\log_2 P(Y=0|x_1=5,x_2=3) = -4/5\cdot 3$

$$\frac{\partial}{\partial w_i}\log_2 P(y=0|x_1,x_2) = -\frac{\partial}{\partial w_i}\log\left(1+2^{w_2x_2+w_1x_1+w_0}\right)$$

$$= -\underbrace{\frac{1}{1+2^{w_2x_2+w_1x_1+w_0}}\cdot 2^{w_2x_2+w_1x_1+w_0}}_{P(y=1|x_1=5,x_2=3)}\cdot\frac{\partial}{\partial w_i}\left(w_2x_2+w_1x_1+w_0\right)$$

**Q4.4 (5 points)** Consider adding an additional input feature, $x_1^2 + x_2^2$.

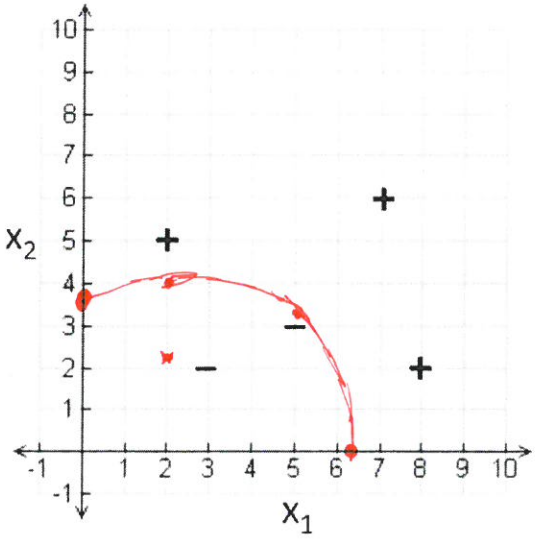
$$P(y = 1|x_1, x_2) = \frac{2^{w_3(x_2^2+x_1^2)+w_2x_2+w_1x_1+w_0}}{1 + 2^{w_3(x_2^2+x_1^2)w_2x_2+w_1x_1+w_0}}. \quad (2)$$

What weights, $w_1, w_2, w_3$, provide a good fit to the data?

+ (2,5): $29w_3 + 5w_2 + 2w_1 + w_0 > 0$
- (3,2): $13w_3 + 2w_2 + 3w_1 + w_0 < 0$
- (5,3): $34w_3 + 3w_2 + 5w_1 + w_0 < 0$
+ (7,6): $75w_3 + 6w_2 + 7w_1 + w_0 > 0$
+ (8,2): $68w_3 + 2w_2 + 8w_1 + w_0 > 0$

$w_3 = 1$
$w_1 = -5$
$w_0 = -10$
$w_2 = 0$

Draw the new decision boundary for this modified logistic regression model on the following plot:



# Q5. Sampling (20 points)

Consider a distribution $p(x)$ that is difficult to directly sample from (or integrate over) and some desired statistic of the distribution, $\mathbb{E}_p[f(X)] = \int_x p(x)f(x)dx = c$.

**Q5.1: (10 points)** Rejection sampling with proposal distribution $q(x)$ (with same support as $p$) and bound $M > 1$ has the following (incomplete) algorithm:

1. Sample x from $q(x)$ and u from $U[0, 1]$
2. Check whether or not ? ? ? ? ? ?
   - If this holds, accept $x$ as a sample from $p(x)$
   - If not, reject the value $x$ and repeat the sampling step.

What is the missing criterion (based on $p(x)$, $q(x)$, $M$ and $u$) for accepting/rejecting the sample in step 2?

$$u < \frac{p(x)}{Mq(x)}$$

What are the requirements on $M$ and why are they needed?

$$Mq(x) \geq p(x) \quad \forall x$$

**Q5.2: (10 points)** A particle filter is often used to generate samples from a hidden state model (like a Hidden Markov Model, defined by $P(S_{t+1}|S_t)$ and $P(O_t|S_t)$) given a set of observations, $o_{1:T}$: $P(S_{1:T}|o_{1:T})$.

Typically, a large number of particles (e.g., 10,000) are employed within a single "run" of the particle filter. If instead we run 10,000 separate particle filters, each with a single particle, what will the sample results be (i.e., how are they distributed)?

They would be samples from

$P(S_{1:T})$ (no conditioning on observations)

If instead we use 5,000 separate particle filters, each with two particles, how will the surviving particles compare to the surviving particles from a single particle filter with 10,000 particles (i.e., how will the distributions differ)?

2 particles gives much higher variance than a single particle filter with 10,000 particles