COMS 4771 (Fall 2017) Final Exam December 07, 2017

Please show all your work! Answers without supporting work will not be given credit.
Write your answers in spaces provided.
There are total 13 pages including two blank pages at the end for scratch work.
You have 1 hour and 15 minutes to complete this exam.
You may use any result from the lectures or the homeworks without proof.

Name & UNI:________________________________________

1. **(15 points)** State True or False. No justification needed!
(+1 for correct, 0 for blank, -1 for incorrect answer)

(a) __________ For any set of $n$ random variables $X_1, \ldots, X_n$, the joint distribution $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | X_1, \ldots, X_{i-1})$.

(b) __________ If two random variables are conditionally independent, then they are not necessarily independent; but if the two random variables are independent then they are necessarily conditionally independent as well.

(c) __________ In a Hidden Markov Model, the observed variable $X_{t-1}$ (at time $t-1$) is conditionally independent of the observed variable $X_{t+1}$ (at time $t+1$) given $X_t$.

(d) __________ Consider the unit $L_p$-ball in $\mathbb{R}^d$, that is the set $\{x \in \mathbb{R}^d : \|x\|_p \leq 1\} =: B_p$. Then, $B_1 \subseteq B_2 \subseteq B_\infty$.

(e) __________ A non-linear kernel transform in sufficiently high dimensions can always achieve zero test error.

(f) __________ The dual variables ($\alpha_i$) in Support Vector Machines (SVMs), take non-zero values only if the corresponding datapoints $x_i$ are in fact "support" vectors that dictate the margin.

(g) __________ VC dimension of decision trees in $\mathbb{R}^2$ is infinite.

(h) __________ The decision boundary induced by a Logistic regression classifier on a two-class problem is always linear.

(i) __________ The Lloyd's algorithm for $k$-means clustering for $k = 2$ can give solutions that are arbitrarily bad in terms of the clustering cost compared to the optimal ($k = 2$) $k$-means solution.

(j) __________ The maximum likelihood setting of the parameters for a mixture model often yields undesirable results.

(k) __________ A directed graphical model on $n$ variables that has a structure of a fully connected directed acyclic graph (DAG) admits no independencies (conditionally or unconditonally) amongst the $n$ variables.

(l) __________ The notation "$A \perp B \mid C$" means "$A$ and $B$ are independent given $C$". Then

$$X \perp Y \mid W, Z \text{ and } X \perp W \mid Y, Z \implies X \perp W, Y \mid Z.$$

(m) __________ $L_2$ (Euclidean) distances can always be computed efficiently in Kernel space via the kernel trick.

(n) __________ For any hypothesis class $\mathcal{F}$, it must be the case that $\mathrm{VC}(\mathcal{F}) \leq \log_2(|\mathcal{F}|)$.

(o) __________ When compared to batch learning, active learning can significantly reduce the number of labelled samples needed to learn a concept to a desired level of accuracy.

2. [**Maximum Likelihood Estimation of fully observed HMMs (15 points)**] Let the distribution of $(X_t, Y_t)_{t\in\{1,2,\ldots\}}$ be from a discrete space HMM, where each $Y_i$ (the hidden state at time $t$) takes in values from $[K] := \{1, 2, \ldots, K\}$ and each $X_t$ (the observation at time $t$) takes values in $[D] := \{1, 2, \ldots, D\}$. Recall that the HMM parameters $\theta = (\pi, A, B)$ have the following semantics: $P[Y_1 = i] = \pi_i$, $P[Y_{t+1} = j|Y_t = i] = A_{i,j}$ and $P[X_t = j|Y_t = i] = B_{i,j}$.

Suppose you have as training data a labeled sequence $((x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T)) \in ([D] \times [K])^T$. What is the maximum likelihood estimation of the HMM parameters $\theta$ given this data? (Hint: if needed, use the convention $0/0 = 0$ in deriving your estimates.)

3. **[Optimization of sparse regression]** Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a data matrix with the $i^{\text{th}}$ row of $\mathbf{X}$ is the $i^{\text{th}}$ data observation $x_i \in \mathbb{R}^d$. Let $y \in \mathbb{R}^n$ be the real-valued labels. Let $\mathcal{W} := \{w \in \mathbb{R}^d :$ $w$ has at most one non-zero entry$\}$. Suppose you want to find the weight vector $\hat{w} \in \mathcal{W}$ that minimizes the objective function $f(w) := \|y - \mathbf{X}w\|^2$ over all $w \in \mathcal{W}$.

   (a) **(5 points)** Formulate this as an optimization problem. (Make sure to clearly indicate the dimension and the variables).

   (b) **(3 points)** Is this a convex optimization problem? (why or why not)?

(c) **(10 points)** Describe an algorithm for computing such a vector $\hat{w}$. Make sure that your pseudocode is clear and precise, and specify what is exactly returned by your algorithm. (Hint: all but one entry of $\hat{w}$ are zero).

(d) **(2 points)** What is the time complexity of your algorithm (give it in terms of the parameters $n$ and $d$)?

Cont.

4. **[Principal Components Analysis (PCA) and beyond]**

Recall from lecture that given a data matrix $X = \begin{bmatrix} | & & | \\ x_1 & \cdots & x_n \\ | & & | \end{bmatrix}$, where each observation $x_i \in \mathbb{R}^d$, the best $k$-dimensional linear mapping that minimizes the squared reconstruction error is given by the eigenvectors corresponding to the top $k$ eigenvalues of the outer product matrix $XX^\mathsf{T}$. This is also called the $k$-dimensional PCA subspace.

(a) Suppose you are informed that the function to compute the eigenvectors and eigenvalues in your favorite language is buggy (so you cannot use this function). As an alternative you explore the language documentation and find a function that has the the ability to decompose any $d \times n$ matrix $X$ into a summation over three sets variables: $\sigma_i$ (a scalar), $u_i$ (a vector in $\mathbb{R}^d$), and $v_i$ (a vector in $\mathbb{R}^n$). That is, the data matrix $X$ can be written as:

$$X = \sum_{i=1}^{d} \sigma_i u_i v_i^\mathsf{T},$$

These variables have special properties:

- The scalars $\sigma_i$'s are all non-negative, such that: $0 \leq \sigma_1 \leq \ldots \leq \sigma_d$.
- The vectors $u_1, \ldots, u_d$ are orthonormal. That is, $\|u_i\| = 1$ and $u_i \cdot u_j = 0$ (for $i \neq j$).
- The vectors $v_1, \ldots, v_d$ are also orthonormal. That is, $\|v_i\| = 1$ and $v_i \cdot v_j = 0$ (for $i \neq j$).

i. **(5 points)** Compute $XX^\mathsf{T}$ in terms of $\sigma_i$, $u_i$ and $v_i$. Simplify your answer as much as possible.

ii. **(5 points)** Using the definition of eigenvector of a matrix $A$: "$Ax = \lambda x$", show that each $u_i$ is an eigenvector of the outerproduct $XX^\mathsf{T}$. What are the corresponding eigenvalues?

iii. **(2 points)** Which $k$ eigenvectors (from $u_1, \dots, u_d$) of $XX^\mathsf{T}$ should be used to form a $k$-dimensional PCA subspace?

(b) **(3 points)** PCA yields a subspace that gives the best reconstruction error, but this subspace can be arbitrarily bad for some predictive tasks such as classification. Suppose we want to do linear classification on binary labelled data in $\mathbb{R}^2$. To reduce representational complexity, we wish to project this data onto a 1D subspace via PCA and perform linear classification in 1D.

Depict an example binary labelled dataset in $\mathbb{R}^2$ such that even though there exists a 1D projection of the dataset that can perfectly linearly separate the two classes, the best linear separator on the 1D PCA projection will give poor classification accuracy. (You must justify your answer why PCA projection would not pick a good label separating subspace by outlining the kinds of properties your depicted dataset must have.)

5. **[Bayes Nets]**

(a) **(5 points)** Draw a Bayes net over the random variables $\{A, B, C, D\}$ where the following conditional independence assumptions hold. Here, $X \perp Y | Z$ means $X$ is conditionally independent of $Y$ given $Z$, and $X \not\perp Y | Z$ means $X$ and $Y$ are not conditionally independent given $Z$, and $\emptyset$ stands for the empty set.

- $A \perp B | \emptyset$
- $A \not\perp D | B$
- $A \perp D | C$
- $A \not\perp C | \emptyset$
- $B \not\perp C | \emptyset$
- $A \not\perp B | D$
- $B \perp D | A, C$

(b) **(2 points)** Write a simple expression for the joint distribution over the variables $\{A, B, C, D\}$ that captures all the conditional (in)dependencies asserted by the Bayes net you depicted in part (a).

$\Pr[A, B, C, D] =$

(c) **(3 points)** List all the conditional independencencies asserted by the Bayes net you depicted in part (a) but with all the arrows reversed.

6. [**A better output Perceptron algorithm guarantee**]
In class, we saw that when the training sample $S$ is linearly separable with a maximum margin $\gamma > 0$, then the Perceptron algorithm run cyclically over $S$ is guaranteed to converge after $T \leq \left(R/\gamma\right)^2$ updates, where $R$ is the radius of the sphere containing the sample points. This does not guarantee however that the hyperplane solution returned by Perceptron, i.e. $w_T$ achieves a margin close to $\gamma$.

   (a) **(5 points)** Show an example training dataset $S$ in $\mathbb{R}^2$ that has margin $\gamma$, and an order of updates made by the Perceptron algorithm where the hyperplane solution returned has arbitrarily bad margin on $S$.

   (b) Consider the following modification to the perceptron algorithm:

**Modified Perceptron Algorithm**

```
Input: training dataset $S = (x_i, y_i)_{i=1,\ldots,n}$
Output: learned vector $w$
- Initialize $w_0 := 0, t := 0$
- while there exists an example $(x, y) \in S$, such that $\overbrace{\cancel{y\,(w_t \cdot x) \leq 0}}^{\text{original condition}} \quad \overbrace{2y(w_t \cdot x) \leq \gamma\|w_t\|}^{\text{modified condition}}$
-     set $w_{t+1} := w_t + yx$
-     set $t := t + 1$
- return $w_t$.
```

      i. **(2 points)** If the Modified Perceptron Algorithm (MPA) terminates after $T$ rounds, what margin guarantee is achieved by the hyperplane $w_T$ returned by MPA? Justify your answer.

ii. **(3 points)** It turns out that for a training sample $S$ of margin $\gamma$ one can also show that at any iteration $t$ where the Modified Perceptron Algorithm (MPA) makes a mistake, the following is true.

A. $w_t \cdot w^* \geq (w_{t-1} \cdot w^*) + \gamma$

B. $\|w_t\| \leq \frac{4R^2}{\gamma} + \frac{3}{4}t\gamma$

From properties A and B, what mistake bound can you derive for MPA? That is, bound the maximum iterations $T$, or equivalently, bound the number of mistakes made by MPA.

**[blank page 1 for scatch work]**

**[blank page 2 for scatch work]**

The End.