CS 771A: Introduction to Machine Learning				**Midsem Exam** (15 Sep 2019)
Name	SAMPLE SOLUTIONS			**80 marks** Page **1** of **6**
Roll No		Dept.		

Instructions:

1. This question paper contains 3 pages (6 sides of paper). Please verify.
2. Write your name, roll number, department in **block letters neatly** with ink **on each page** of this question paper.
3. If you don't write your name and roll number on **all** pages, **pages may get lost** when we unstaple to scan pages
4. Write your final answers neatly **with a blue/black pen**. Pencil marks may get smudged.
5. Don't overwrite/scratch answers especially in MCQ and T/F. We will entertain no requests for leniency.

Q1. Write **T** or **F** for True/False (write **only in the box on the right hand side**) **(10x2=20 marks)**

1	When using kNN to do classification, using a large value of k always gives better performance since more training points are used to decide label of the test point	F
2	Cross validation means taking a small subset of the test data and using it to get an estimate of how well will our algorithm perform on the entire test dataset	F
3	The EM algo does not require a careful initialization of model parameters since it anyway considers all possible assignments of latent variables with different weights	F
4	If X and Y are two real-valued random variables such that $\text{Cov}(X,Y) < 0$ then at least one of X or Y must have negative variance i.e. either $\mathbb{V}X < 0$ or $\mathbb{V}Y < 0$	F
5	If $\mathbf{a} \in \mathbb{R}^2$ is a constant vector and $f: \mathbb{R}^2 \to \mathbb{R}$ is such that $g(\mathbf{x}) = f(\mathbf{x}) + \mathbf{a}^\top\mathbf{x}$ is a non-convex function, then $h(\mathbf{x}) = f(\mathbf{x}) - \mathbf{a}^\top\mathbf{x}$ must be a non-convex function too	T
6	The SVM is so named because the decision boundary of the SVM classifier passes through the data points which are marked as being support vectors	F
7	Suppose X is a real valued random variable with variance $\mathbb{V}X = 9$. Then the random variable Y defined as $Y = X - 2$ will always satisfy $\mathbb{V}Y = \mathbb{V}X - 2^2 = 5$	F
8	The LwP algorithm for binary classification always gives linear decision boundary if we use one prototype per class and Euclidean distance to measure distances	T
9	If $f, g: \mathbb{R}^2 \to \mathbb{R}$ are two non-convex functions, then the function $h: \mathbb{R}^2 \to \mathbb{R}$ defined as $h(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$ must always be non-convex too	F
10	If we learn models $\{\mathbf{w}^c\}_{c=1}^C$ for multiclassification using the Crammer-Singer loss function, these models can be used to assign a PMF over the class labels $[C]$	T

Q2 Phase retrieval is used in X-ray crystallography. Let $\mathbf{x}^i \in \mathbb{R}^d, i \in [n]$ be features and $y^i \in \mathbb{R}$ be labels. All data points are independent. However, we only get to see the absolute value of labels, i.e. the train data is $\left\{\left(\mathbf{x}^i, u^i\right)\right\}_{i=1}^n$ where $u^i = \left|y^i\right|$. Let $z^i \in \{-1,1\}$ be latent variables for missing label signs (aka *phases*). Use the data likelihood function $\mathbb{P}\left[u^i \mid z^i, \mathbf{x}^i, \mathbf{w}\right] = \mathcal{N}\left(u^i z^i\, ;\, \mathbf{w}^\top\mathbf{x}^i, 1\right)$. Note that this is a discriminative setting (i.e. $\mathbf{x}^i$ are constants). Expressions in your answers may contain unspecified normalization constants. Give only brief derivations. **(8+6+6=20 marks)**

2.1 Assuming $\mathbb{P}\left[z^i = c \mid \mathbf{x}^i, \mathbf{w}\right] = \mathbb{P}\left[z^i = c\right] = 0.5$ for $c \in \{-1,1\}$ (i.e. uniform prior on z^i that does not depend on features or model), derive an expression for $\mathbb{P}\left[z^i = 1 \mid u^i, \mathbf{x}^i, \mathbf{w}\right]$. Using this, derive an expression for the MAP estimate $\underset{c\in\{-1,+1\}}{\arg\max}\ \mathbb{P}\left[z^i = c \mid u^i, \mathbf{x}^i, \mathbf{w}\right]$

Applying Bayes rule and $\mathbb{P}\left[z^i = 1 \mid \mathbf{x}^i, \mathbf{w}\right] = 0.5$, we have (omitting normalization constants)

$$\mathbb{P}\left[z^i = 1 \mid u^i, \mathbf{x}^i, \mathbf{w}\right] = \frac{\mathbb{P}\left[u^i \mid z^i = 1, \mathbf{x}^i, \mathbf{w}\right] \cdot \mathbb{P}\left[z^i = 1 \mid \mathbf{x}^i, \mathbf{w}\right]}{\mathbb{P}\left[u^i \mid \mathbf{x}^i, \mathbf{w}\right]} \propto \exp\left(-\frac{\left(u^i - \mathbf{w}^\top \mathbf{x}^i\right)^2}{2}\right)$$

Similarly, we have $\mathbb{P}\left[z^i = -1 \mid u^i, \mathbf{x}^i, \mathbf{w}\right] \propto \exp\left(-\frac{\left(-u^i - \mathbf{w}^\top \mathbf{x}^i\right)^2}{2}\right)$. This tells us that we should set z^i to whatever value that leads to a smaller residual error. A nice way of saying this is

$$\underset{c \in \{-1,+1\}}{\arg\max}\, \mathbb{P}\left[z^i = c \mid u^i, \mathbf{x}^i, \mathbf{w}\right] = \operatorname{sign}\left(\left|-u^i - \mathbf{w}^\top \mathbf{x}^i\right| - \left|u^i - \mathbf{w}^\top \mathbf{x}^i\right|\right)$$

where we break ties (when both terms on the RHS are equal) arbitrarily, say in favour of 1. We may choose to break ties any way we wish since $\operatorname{sign}(0)$ is not cleanly defined and does not matter in calculations.

2.2 Derive an expression for $\mathbb{P}[\mathbf{w} \mid \mathbf{u}, \mathbf{z}, X]$ using a standard Gaussian prior $\mathbb{P}[\mathbf{w}] = \mathcal{N}(\mathbf{0}, I_d)$. Then derive an expression for the MAP estimate for $\mathbf{w}$ i.e. $\underset{\mathbf{w} \in \mathbb{R}^d}{\arg\max}\, \mathbb{P}[\mathbf{w} \mid \mathbf{u}, \mathbf{z}, X]$ (here we are using shorthand notation $X = [\mathbf{x}^1, \ldots \mathbf{x}^n]^\top \in \mathbb{R}^{n \times d}, \mathbf{u} = [u^1, \ldots, u^n] \in \mathbb{R}^n, \mathbf{z} = [z^1, \ldots, z^n] \in \mathbb{R}^n$).

Using independence, the Bayes rule, and ignoring proportionality constants as before gives us

$$\mathbb{P}[\mathbf{w} \mid \mathbf{u}, \mathbf{z}, X] \propto \mathbb{P}[\mathbf{u} \mid \mathbf{w}, \mathbf{z}, X] \cdot \mathbb{P}[\mathbf{w}] \propto \exp\left(-\frac{1}{2}\|\mathbf{w}\|_2^2\right) \cdot \prod_{i=1}^n \exp\left(-\frac{\left(u^i z^i - \mathbf{w}^\top \mathbf{x}^i\right)^2}{2}\right)$$

Note that the expression for $\mathbb{P}\left[u^i \mid z^i, \mathbf{x}^i, \mathbf{w}\right]$ is available to us from the question text itself. Taking logarithms as usual gives us

$$\hat{\mathbf{w}}_{\text{MAP}} = \underset{\mathbf{w} \in \mathbb{R}^d}{\arg\min}\, \|\mathbf{w}\|_2^2 + \sum_{i=1}^n \left(u^i z^i - \mathbf{w}^\top \mathbf{x}^i\right)^2$$

Applying first order optimality and using the shorthand $v^i = u^i z^i$ and $\mathbf{v} = [v^1, \ldots, v^n] \in \mathbb{R}^n$

$$\hat{\mathbf{w}}_{\text{MAP}} = (X^\top X + I_d)^{-1} X^\top \mathbf{v}$$

2.3 Using the above derivations, give the pseudocode (as we write in lecture slides i.e. not necessarily Python code or C code but sufficient details of the algorithm updates) for an alternating optimization algorithm for estimating the model $\mathbf{w}$ in the presence of the latent variables. Give precise update expressions in your pseudocode and not just vague statements.

```
AltOpt for Phase Retrieval
1. Initialize model $\mathbf{w}$
2. For $i \in [n]$, update $\{z_i\}$ using $\{\mathbf{w}^c\}$
    1. Let $z_i = \text{sign}\left(\left|-u^i - \mathbf{w}^\top \mathbf{x}^i\right| - \left|u^i - \mathbf{w}^\top \mathbf{x}^i\right|\right)$
    2. Break ties arbitrarily
3. Update $\mathbf{w}$ using $\{z_i\}$
    1. Let $v^i = u^i z^i$ and $\mathbf{v} = [v^1, \ldots, v^n]$
    2. Let $\mathbf{w} = (X^\top X + I_d)^{-1} X^\top \mathbf{v}$
4. Repeat until convergence
```

Q3 We have seen that algorithms such as the EM require weighted optimization problems to be solved where different data points may have different weights. Consider the following problem of L2 regularized squared hinge loss minimization but with different weights per data point. The data points are $\mathbf{x}^i \in \mathbb{R}^d$ and the labels are $y^i \in \{-1,1\}$. The weights q_i are all known (i.e. are constants) and are all strictly positive i.e. $q_i > 0, q_i \neq 0$ for all $i = 1, \ldots, n$ **(3+2+5=10 marks)**

$$\arg\min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} q_i \cdot \left(\left[1 - y^i \cdot \mathbf{w}^\top \mathbf{x}^i\right]_+\right)^2$$

3.1 As we did in assignment 1, rewrite the above problem as an equivalent problem that has inequality constraints in it (the above problem does not have any constraints).

$$\arg\min_{\substack{\mathbf{w} \in \mathbb{R}^d \\ \boldsymbol{\xi} \in \mathbb{R}^n}} \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} q_i \cdot \xi_i^2$$

$$\text{s.t. } y^i \cdot \mathbf{w}^\top \mathbf{x}^i \geq 1 - \xi_i, \text{ for all } i \in [n]$$

Similar to what we observed in assignment 1, even in this case, including or omitting the constraints $\xi_i \geq 0$ does not affect the solution.

3.2 Then introduce dual variables as appropriate and write down the expression for the dual problem as a max-min problem (no need to write the Lagrangian expression separately).

The Lagrangian is $\mathcal{L}(\mathbf{w},\boldsymbol{\xi},\boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^n q_i \cdot \xi_i^2 + \sum_{i=1}^n \alpha_i\left(1 - \xi_i - y^i \cdot \mathbf{w}^\top \mathbf{x}^i\right)$

Thus, the dual problem is

$$\max_{\boldsymbol{\alpha}\geq 0} \left\{ \min_{\mathbf{w},\boldsymbol{\xi}} \left\{ \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^n q_i \cdot \xi_i^2 + \sum_{i=1}^n \alpha_i\left(1 - \xi_i - y^i \cdot \mathbf{w}^\top \mathbf{x}^i\right) \right\} \right\}$$

3.3 Simplify the dual by eliminating the primal variables and write down the expression for the simplified dual. Show only brief derivations.

Applying first order optimality to the inner unconstrained optimization problem gives us:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^n \alpha_i\, y^i \cdot \mathbf{x}^i$$

$$\frac{\partial \mathcal{L}}{\partial \xi} = 0 \Rightarrow \xi_i = \frac{\alpha_i}{2q_i}$$

Putting these in the dual expression gives us the following simplified dual problem

$$\max_{\boldsymbol{\alpha}\geq 0} \left\{ \boldsymbol{\alpha}^\top \mathbf{1} - \frac{1}{2}\boldsymbol{\alpha}^\top (Q + D)\boldsymbol{\alpha} \right\}$$

where Q is an $n \times n$ matrix with $Q_{ij} = \alpha_i \alpha_j y^i y^j \langle \mathbf{x}^i, \mathbf{x}^j \rangle$ and D is an $n \times n$ diagonal matrix with $D_{ii} = \frac{1}{2q_i}$ and $D_{ij} = 0$ if $i \neq j$.

Q4 Recall the uniform distribution over an interval $[a, b] \subset \mathbb{R}$ where $a < b$. Just two parameters, namely a, b, are required to define this distribution (no restrictions on a, b being positive/non-zero etc, just that we must have $a < b$. Note this implies $a \neq b$). The PDF of this distribution is

$$\mathbb{P}[x \mid a, b] = \mathcal{U}(x; a, b) \triangleq \begin{cases} 0 & x < a \\ 1/(b-a) & x \in [a, b] \\ 0 & x > b \end{cases}$$

Given n independent samples $x^1, \dots, x^n \in \mathbb{R}$ (assume w.l.o.g. that not all samples are the same number) we wish to learn a uniform distribution as a generative distribution using these samples using the MLE technique i.e. we wish to find

$$\underset{a<b, a\neq b}{\arg\max}\ \mathbb{P}[x^1, \dots, x^n \mid a, b]$$

Give a brief derivation for, and the final values of, $\hat{a}_{\mathrm{MLE}}$ and $\hat{b}_{\mathrm{MLE}}$. **(5+5=10 marks)**

Using independence, we have $\underset{a<b,a\neq b}{\arg\max}\ \mathbb{P}[x^1,\ldots,x^n \mid a,b] = \underset{a<b,a\neq b}{\arg\max}\ \prod_{i=1}^n \mathcal{U}(x^i;a,b)$

Now, suppose we have a pair (a,b) such that for some $i \in [n]$, we have $x^i \neq [a,b]$, then $\mathcal{U}(x^i;a,b) = 0$ and as a result $\mathbb{P}[x^1,\ldots,x^n \mid a,b] = 0$ too! This means that if we denote $m \triangleq \min_i x^i$ and $M \triangleq \max_i x^i$, then we must have $a \leq m$ and $b \geq M$ to get a non-zero value of the likelihood function i.e. we need to solve

$$\underset{a\leq m,b\geq M}{\arg\max} \prod_{i=1}^n \mathcal{U}(x^i;a,b) = \underset{a\leq m,b\geq M}{\arg\max} \left(\frac{1}{b-a}\right)^n$$

The above is maximized for the smallest value of $b-a$ which, subject to the constraints, is achieved exactly at $a=m, b=M$. Thus, we have

$$\hat{a}_{\text{MLE}} = \min_i x^i \text{ and } \hat{b}_{\text{MLE}} = \max_i x^i$$

Q5. Fill the circle (**don't tick**) next to all the correct options (**many may be correct**).(**2x3=6 marks**)

5.1 The use of the Laplace (aka Laplacian) prior and Laplace (aka Laplacian) likelihood results in a MAP problem that requires us to solve an optimization problem whose objective function is

A	Always convex and always differentiable	○
B	Always convex but possibly non-differentiable	●
C	Possibly non-convex but always differentiable	○
D	Always non-convex and always non-differentiable	○

5.2 In probabilistic multiclassification with C classes, if for a test data point, the ML algorithm predicts a PMF over the classes with an extremely small variance, then it means that

A	The mode of that PMF should have a probability value much larger than 0	●
B	The mode of that PMF should have a probability value very close to 0	○
C	The ML algorithm is very confident about its prediction on that data point	●
D	The ML algorithm is very unsure about its prediction on that data point	○

Q6 Nadal and Federer have played a total of 80 matches of which Nadal won 50, Federer won 30. They have played on three types of courts – clay, grass, and hard. Among the matches Nadal won, 70% were played on clay courts, 4% on grass courts and rest on hard courts. Federer has won a 15/120 fraction of matches played on clay courts, 96/120 fraction of matches played on grass courts, and 68/120 fraction of matches played on hard courts. What is the **number of matches** that the two players have played on each of the three types of courts? **(3x2=6 marks)**

Clay (40) Grass (10) Hard (30)

Q7 Let X be a discrete random variable with support $\{-1,0,1\}$. Find a PMF for X for which X has the highest possible variance. What value of variance do you get in this case? Repeat the analysis (i.e. give the highest variance PMF as well as the variance value) when X is a Rademacher random variable i.e. has support only over $\{-1,1\}$. Justify all your answers briefly. **(3+1+3+1=8 marks)**

Suppose the PMF assigns probability values p_{-1}, p_0, p_1 to the support elements. Then we have $\mathbb{E}X = (p_1 - p_{-1})$ and $\mathbb{V}X = \mathbb{E}[X^2] - (\mathbb{E}X)^2 = (p_1 + p_{-1}) - (p_1 - p_{-1})^2$. Now, whereas we could go all Lagrangian on this problem and solve it by brute force, a more careful look at the problem gives results more readily.

The largest (perhaps unachievably so) value of the last expression is achieved when $p_1 + p_{-1}$ takes on its largest value (which is 1 since $p_{-1} + p_0 + p_1 = 1$ and $p_0 \geq 0$) and $(p_1 - p_{-1})^2$ takes on its smallest value (which is 0 since a square of a real number can never be negative). Thus, we must not expect a result better than $\mathbb{V}X = 1$.

However, the above can actually be achieved. $(p_1 - p_{-1})^2 = 0$ when $p_1 = p_{-1}$ and we cab even simultaneously ensure $p_1 + p_{-1}$ by setting $p_1 = p_{-1} = 0.5$. Thus, at the PMF $\{p_{-1} = 0.5, p_0 = 0, p_1 = 0.5\}$, the random variable has highest variance of $\mathbb{V}X = 1$.

For the Rademacher case, the solution is readily seen to be $\{p_{-1} = 0.5, p_1 = 0.5\}$ and $\mathbb{V}X = 1$ since in the first case, the solution we obtained looks exactly like a Rademacher random variable if we look at the support elements which are assigned non-zero probability.

- END OF EXAM -