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	
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	
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	
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 $\mathbb{V}Y < 0$	
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	
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	
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 $\mathbb{V}Y = \mathbb{V}X - 2^2 = 5$	
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	
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	
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 $[C]$	

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

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. $\arg\max_{\mathbf{w}\in\mathbb{R}^d} \mathbb{P}[\mathbf{w} \mid \mathbf{u}, \mathbf{z}, X]$ (here we are using shorthand notation $X = [\mathbf{x}^1, \dots \mathbf{x}^n]^\top \in \mathbb{R}^{n\times d}, \mathbf{u} = [u^1, \dots, u^n] \in \mathbb{R}^n, \mathbf{z} = [z^1, \dots, z^n] \in \mathbb{R}^n$).

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.

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).

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).

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

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 $x^1, \ldots, x^n \in \mathbb{R}$ (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

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

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

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 (   ) Grass (   ) Hard (   )

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)**

------------------------------- END OF EXAM -------------------------------