Name:

Roll No.:

Dept.:

IIT Kanpur
CS771 Intro to ML
Mid-semester Examination
Date: September 20, 2018

Instructions:

Total: **80 marks**

1. This question paper contains a total of 8 pages (8 sides of paper). Please verify.
2. Please write your name, roll number, department on **every side of every sheet** of this booklet.
3. Write final answers **neatly with a pen**. Pencil marks can get smudged and you may lose credit.
4. **Important:** Please do not give derivations/elaborate steps unless specifically asked for it. Feel free to use standard results (e.g., solution of least squares regression) without deriving them from scratch. If needed, you may use the personal rough space for more detailed derivations.
5. The last page of the question paper lists some formulae if you need them.

Section 1 (True or False: 10 X 1 = 10 marks). For each of the following simply write **T** or **F** in the box.

1. The prediction cost (time taken to predict the label for a test input) for 1-nearest neighbors classification is higher as compared to that of prototype based classification.
2. A least squares regression problem with ℓ_1 norm regularizer on the weight vector $\boldsymbol{w}$ will have a globally optimal solution.
3. The softmax regression based discriminative model for classification with K classes would require learning $K-1$ probability distributions.
4. Changing the Perceptron update rule from $\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} + y_n \boldsymbol{x}_n$ to $\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} + \gamma y_n \boldsymbol{x}_n$ would effectively learn the same hyperplane separator.
5. The size of kernel induced feature mapping ϕ of a polynomial kernel with degree $d \geq 2$ is the same for any value of d.
6. If the MAP objective has a unique optima then the predictive distribution computed using the MAP estimate and computed by averaging over the full posterior will be the same.
7. For linear/logistic regression, it is not possible to regularize different entries of the weight vector differently if using a zero mean Gaussian prior over the weight vector.
8. Increasing the parameter C in soft-margin SVM objective $\frac{||\boldsymbol{w}||^2}{2} + C\sum_{n=1}^{N} \xi_n$ tends to increase the ℓ_2 squared norm of $\boldsymbol{w}$.
9. Running a linear model on landmark based features or kernel random features would always be faster than a linear model on the original features.
10. The SGD algorithm for a binary linear classification model would update the weight vector only when the current weight vector mispredicts the chosen training example.

Section 2 (8 problems: 8 x 3 = 24 marks). Write your answers precisely and concisely in the provided box.

1. Consider a data set with 5 points $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4, \boldsymbol{x}_5\}$ in two dimensions: $\{(1,0), (-1,0), (0,1), (3,0), (3,1)\}$. Run two iterations of K-means with initial points at $\mu_1 = (-1,0)$ and $\mu_2 = (3,1)$. What are the assignments z_1, z_2, z_3, z_4, z_5 and the centers μ_1 and μ_2 at each iteration? Has the algorithm converged?

Name:

Roll No.: **Dept.**:

2. Rank the following methods in terms of the test-time prediction cost: linear SVM, kernelized Perceptron, and kernelized SVM (fastest first, slowest last), and briefly justify your answer.

3. Consider multi-label classification given training data $\{(\boldsymbol{x}_n \boldsymbol{y}_n)\}_{n=1}^N$. Here each output $\boldsymbol{y}_n \in \{0,1\}^L$, i.e., a binary vector of length L. Briefly describe an approach to learn a multi-label classification model using this data if you only have access to a *binary classification* algorithm $\mathcal{B}$ (e.g., a binary SVM).

4. Consider learning a decision tree, given some training data where each input has D binary-valued features. Let's assume that we will not test any feature that has been tested at one of the previous levels (but we can possibly test a feature at multiple nodes at the same level). How many information gain calculations would be needed to construct the full decision tree (i.e., assuming no pruning)? Just give the basic expression; no need to try simplifying it too much to get a more compact expression.

5. Consider a binary classification dataset with two-dimensional inputs, where each input is of the form $\boldsymbol{x} = (x_1, x_2)$. Suppose the decision boundary is given by $\frac{(x_1-1)^2}{2} + \frac{(x_2-2)^2}{3} = 1$. Note that this is a nonlinear boundary (equation of an ellipse). Write down a mapping $\phi(\boldsymbol{x})$ that will make it possible to separate the two classes using a linear separator. Also write down the weights of this linear model.

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS771 Intro to ML
Mid-semester Examination
Date: September 20, 2018

6. Consider solving multi-class classification by reducing it to binary SVM classification problems using One-vs-All (OvA) and All-Pairs (note: all-pairs is also called "All vs All" or AvA). Suppose we have K classes and M examples per class ($N = KM$ examples total). Typical binary SVM takes roughly N^2 time to train on N examples. For what values of K is it computationally cheaper to use OVA instead of AVA?

7. Consider the hard-margin versions of the SVDD problem (left) and the one-class SVM problem (right).

$$\min_{\boldsymbol{c}\in\mathbb{R}^D, R\in\mathbb{R}} R^2 \qquad\qquad \min_{\boldsymbol{c}\in\mathbb{R}^D, \tau\in\mathbb{R}} \frac{||\boldsymbol{c}||^2}{2} - \tau$$

$$\text{subject to } ||\boldsymbol{x}_n - \boldsymbol{c}||^2 \leq R^2 \qquad \text{subject to } \boldsymbol{c}^\top \boldsymbol{x}_n \geq \tau + \frac{||\boldsymbol{x}_n||^2}{2}$$

Show that these two problems are equivalent for some value of τ. What is that value τ?

8. Consider a least squares regression problem with N examples $(\mathbf{X}, \boldsymbol{y}) = \{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ and regression weight vector $\boldsymbol{w} \in \mathbb{R}^D$. Assume no regularization on $\boldsymbol{w}$. However, suppose we add another M "fake" examples $(\tilde{\mathbf{X}}, \tilde{\boldsymbol{y}}) = \{(\tilde{\boldsymbol{x}}_m, \tilde{y}_m)\}_{m=1}^M$. Show that using these additional fake examples is equivalent to using an ℓ_2 regularizer. For what values of $(\tilde{\mathbf{X}}, \tilde{\boldsymbol{y}})$, it will correspond to an ℓ_2 regularizer $\lambda \boldsymbol{w}^\top \boldsymbol{w}$?

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS771 Intro to ML
Mid-semester Examination
Date: September 20, 2018

Section 3 (6 problems: 6 x 6 = 36 marks). Write your answers precisely and concisely in the provided box.

1. Assuming hard-margin SVM, show that, given the solution for the dual variables α_n's, the bias term $b \in \mathbb{R}$ can be computed as $b = y_s - t_s$ where s can denote the index of *any* of the support vectors, and t_s is a term that requires computing a summation defined over all the support vectors. (Hint: Use KKT conditions)

2. Show that we can rewrite regression with *absolute* loss function $|y_n - \boldsymbol{w}^\top \boldsymbol{x}_n|$ as a *reweighted* least squares objective where the squared loss term for each example $(\boldsymbol{x}_n, y_n)$ is multiplied by an importance weight $s_n > 0$. Write down the expression for s_n, and briefly explain why this expression for s_n makes intuitive sense. Given N examples $\{(\boldsymbol{x}_n, y_n\}_{n=1}^N$, briefly outline the steps of an optimization algorithm that estimates the unknowns ($\boldsymbol{w}$ and the importance weights $\{s_n\}_{n=1}^N$) for this reweighted least squares problem.

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS771 Intro to ML
Mid-semester Examination
Date: September 20, 2018

3. Consider the following way to generate a binary random number y: Draw K count-valued random variables $m_k \sim \text{Poisson}(\lambda_k)$, $k = 1, \ldots, K$. Define $m = \sum_{k=1}^{K} m_k$, and generate y as $y = \mathbb{I}[m > 0]$, where $\mathbb{I}[.]$ is indicator function. Derive the expression for $p(y = 1|\lambda_1, \ldots, \lambda_K)$.

4. Suppose we wish to do an "online" K-means by performing an SGD-style optimization on the K-means objective $\sum_{n=1}^{N} ||\boldsymbol{x}_n - \mu_{z_n}||^2 = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk}||\boldsymbol{x}_n - \mu_k||^2$. Assume mini-batch size = 1, i.e., you get one randomly chosen example $\boldsymbol{x}_n$. Assume learning rate η. What will be the SGD update equations for the cluster means $\mu_1, \ldots, \mu_K$? In which direction does each mean move as a result of these SGD updates?

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS771 Intro to ML
Mid-semester Examination
Date: September 20, 2018

5. Consider N scalar-valued observations $x_1, \ldots, x_N$ from a Gaussian $\mathcal{N}(\mu, \lambda^{-1})$. Suppose the mean μ is known and precision λ is unknown. Assume a gamma prior on λ, i.e., $p(\lambda) = \text{Gamma}(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$. Compute the posterior distribution of λ, i.e., $p(\lambda|x_1, \ldots, x_N)$. Is the posterior available in closed form? If yes, why, and what's the name of this distribution? If no, why not? (Hint/suggestion: Try computing the posterior first, before answering yes or no).

6. Consider linear regression with squared loss $\mathcal{L}(\boldsymbol{w}) = \sum_{n=1}^{N} (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2$ (no regularizer), and apply Newton's method to find the optimal $\boldsymbol{w}$. Write down the expression for the weight update at each iteration. What is the minimum number of iterations that Newton's method will take to converge? Is that what you would expect Newton's method to do for *this* problem? If yes, why? If no, why not?

Name:

Roll No.: **Dept.:**

IIT Kanpur
CS771 Intro to ML
Mid-semester Examination
Date: September 20, 2018

Section 4 (1 problem: 10 marks). Write your answers precisely and concisely in the provided box.

1. The mistake-driven Perceptron update rule, after Percentron makes a mistake on $(\boldsymbol{x}_n, y_n)$, updates the weight vector as $\boldsymbol{w}^{(t)} = \boldsymbol{w}^{(t-1)} + y_n\boldsymbol{x}_n$. This update will modify the weight vector (hyperplane separator) in such a way that it becomes *less incorrect* on the example $(\boldsymbol{x}_n, y_n)$, but not necessarily *correct*. For example, if $y_n = 1$ then ${\boldsymbol{w}^{(t)}}^\top\boldsymbol{x}_n$ will become less negative than ${\boldsymbol{w}^{(t-1)}}^\top\boldsymbol{x}_n$ (but not necessarily positive). Likewise, if $y_n = -1$ then ${\boldsymbol{w}^{(t)}}^\top\boldsymbol{x}_n$ will become less positive than ${\boldsymbol{w}^{(t-1)}}^\top\boldsymbol{x}_n$ (but not necessarily negative).

   Let's try to design a variant of Perceptron that guarantees that, after the update, the new weight vector $\boldsymbol{w}^{(t)}$ will *definitely* be correct on the example $(\boldsymbol{x}_n, y_n)$, i.e., $y_n{\boldsymbol{w}^{(t)}}^\top\boldsymbol{x}_n \geq 0$. At the same time, let's not make the new weight vector $\boldsymbol{w}^{(t)}$ drift too far from the current weight $\boldsymbol{w}^{(t-1)}$, i.e., we want to keep the squared ℓ_2 distance $||\boldsymbol{w}^{(t)} - \boldsymbol{w}^{(t-1)}||^2$ as small as possible.

   Formulate the above as an optimization problem and solve it to derive the updates for this variant of the Perceptron. Also verify that your obtained expression for $\boldsymbol{w}^{(t)}$ does satisfy the constraint $y_n{\boldsymbol{w}^{(t)}}^\top\boldsymbol{x}_n \geq 0$.

(if needed, you may continue the answer in the box on the next page)

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS771 Intro to ML
Mid-semester Examination
Date: September 20, 2018

Some formulae you might need

- Gaussian PDF: $\mathcal{N}(x|\mu, \lambda^{-1}) = \sqrt{\frac{\lambda}{2\pi}} \exp(-\frac{\lambda}{2}(x-\mu)^2)$
- Poisson PMF: $\text{Poisson}(x|\lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$
- Hessian of a scalar-valued function $\mathcal{L}(\boldsymbol{w})$ w.r.t. $\boldsymbol{w} \in \mathbb{R}^D$ is $\mathbf{H} = \frac{\partial^2 \mathcal{L}(\boldsymbol{w})}{\partial \boldsymbol{w} \partial \boldsymbol{w}^\top} = \frac{\partial}{\partial \boldsymbol{w}} \left[\frac{\partial \mathcal{L}(\boldsymbol{w})}{\partial \boldsymbol{w}} \right]^\top = \frac{\partial \boldsymbol{g}^\top}{\partial \boldsymbol{w}}$
- Derivatives - linear form: $\frac{\partial \boldsymbol{x}^\top \boldsymbol{a}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{a}^\top \boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a}$, quadratic form: $\frac{\partial}{\partial \boldsymbol{x}} (\boldsymbol{x} - \boldsymbol{s})^\top \mathbf{W} (\boldsymbol{x} - \boldsymbol{s}) = 2\mathbf{W}(\boldsymbol{x} - \boldsymbol{s})$