Name:

Roll No.: **Dept.**:

IIT Kanpur
CS771 Intro to ML
End-semester Examination
Date: November 17, 2017

Instructions: *Total:* **120 marks**

1. This question paper contains a total of 8 pages (8 sides of paper). Please verify.
2. 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. Do not give derivations/elaborate steps unless the question specifically asks you to provide these.

Problem 1 (True or False: 12 X 1 = 12 marks). For each of the following simply write **T** or **F** in the box.

1. The time it takes to make a prediction using a decision tree depends on the number of nodes in that tree.
2. If $f(\mathbf{x})$ is a convex function for $\mathbf{x} \in \mathbb{R}^d$ and $g(\mathbf{x}) = \langle \mathbf{v}, \mathbf{x} \rangle + c$ for some fixed vector $\mathbf{v} \in \mathbb{R}^d$ and $c \in \mathbb{R}$, then $f + g$ is always a convex function.
3. The k-means++ algorithm for clustering, initializes the cluster centers to k points in the dataset that are closest to each other.
4. In CNNs, a larger pool size, e.g., max pooling a larger number of neurons together in a single pool, preserves more information about the output of the layer to which pooling is applied.
5. When working with large datasets, held-out validation is cheaper to execute as compared to k-fold cross validation.
6. The k-means++ algorithm cannot be used when performing kernel k-means clustering with a nonlinear Mercer kernel with an infinite dimensional feature map.
7. If we learn a single model from a model class and find that the learnt model is overfitting to the data, then between bagging and boosting, boosting is better way to fix the problem.
8. The Power method can be used to solve the PCA problem but it cannot be used to solve the kernel PCA problem.
9. Solving the SVM problem is cheaper when using a linear kernel than it is when using the Gaussian kernel.
10. A neural network with a single hidden layer and a single output node with all nodes except input layer nodes using the sigmoid activation function will always learn a continuous function.
11. For small scale recommendation problems, say with only 10 items to recommend from, we can cast the problem as 10 separate classification problems.
12. When interacting with a typical recommendation system, users usually tell the recommendation system what items they like and what items they do not like.

Problem 2 (Ultra Short Answer: 6 x 4 = 24 marks). Give your answers in the space provided only.

1. Write down below, a feature map corresponding to the Mercer kernel $K(\mathbf{z}^1, \mathbf{z}^2) = (\langle \mathbf{z}^1, \mathbf{z}^2 \rangle)^2$ where $\mathbf{z}^i = (x_i, y_i), i = 1, 2$ are 2D vectors. Note that maps will smaller dimensionality will get more credit.

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS771 Intro to ML
End-semester Examination
Date: November 17, 2017

2. I have 1000 data points which I wish to split into a training and a held-out validation set. Tom tells me to take 990 points as training and 10 as validation. Dick declares that dividing into 700 training points and 300 validation points is prefereable whereas Harry has heard that taking 10 training and 990 validation points works best. Which friend should I agree with? Why? Why should I disagree with the other two?

3. My friend has trained a binary classifier which gets only 10% classification accuracy. What is the simplest thing I can do to boost the accuracy of this classifier to a more respectable level?

4. Let K_{int} be the intersection kernel: for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, $K_{\text{int}}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{d} \min\left\{\mathbf{x}_i, \mathbf{y}_i\right\}$. Let $\phi_{\text{int}} : \mathbb{R}^d \to \mathbb{R}^D$ be a feature map corresponding to K_{int}. Write down the expression for $\left\|\phi_{\text{int}}(\mathbf{x}) - \phi_{\text{int}}(\mathbf{y})\right\|_2^2$.

5. I have a regression dataset $\left\{(\mathbf{x}^i, y^i)\right\}_{i\in[n]}$, $\mathbf{x}^i \in \mathcal{X}, y^i \in \mathbb{R}$ and a kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Let $G \in \mathbb{R}^{n\times n}$ denote the Gram matrix with $G_{ij} = K(\mathbf{x}^i, \mathbf{x}^j)$. I perform landmarking with all training points as landmarks i.e. $\hat{\phi}(\mathbf{x}) = [K(\mathbf{x}, \mathbf{x}^1), \ldots, K(\mathbf{x}, \mathbf{x}^n)] \in \mathbb{R}^n$. Solve $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathbb{R}^n} \lambda \cdot \|\mathbf{w}\|_2^2 + \sum_{i=1}^{n} \left(y^i - \left\langle \mathbf{w}, \hat{\phi}(\mathbf{x}^i)\right\rangle\right)^2$ (i.e. ridge regression using the landmarked feature map $\hat{\phi}$) and write down the expression for $\hat{\mathbf{w}}$.

Name:

Roll No.: **Dept.**:

IIT Kanpur
CS771 Intro to ML
End-semester Examination
Date: November 17, 2017

6. Note that the predictor we learnt in part 5 looks like $\left\langle \hat{\mathbf{w}}, \hat{\phi}(\mathbf{x}) \right\rangle = \sum_{i=1}^n \gamma_i \cdot K(\mathbf{x}, \mathbf{x}^i)$ where $\gamma_i = \hat{\mathbf{w}}_i$. Now let $\phi_K : \mathcal{X} \to \mathcal{H}$ be a feature map for the kernel K so that $K(\mathbf{x}^i, \mathbf{x}^j) = \langle \phi_K(\mathbf{x}^i), \phi_K(\mathbf{x}^j) \rangle$ for all $\mathbf{x}^i, \mathbf{x}^j \in \mathcal{X}$. Suppose we had instead solved $\hat{\mathbf{W}} = \arg\min_{\mathbf{W} \in \mathcal{H}} \lambda \cdot \|\mathbf{W}\|_{\mathcal{H}}^2 + \sum_{i=1}^n \left(y^i - \langle \mathbf{W}, \phi_K(\mathbf{x}^i) \rangle\right)^2$, i.e. performed kernel ridge regression on the dataset directly instead of landmarking then, as we saw in class, we would have obtained a predictor $\left\langle \hat{\mathbf{W}}, \phi_K(\mathbf{x}) \right\rangle = \sum_{i=1}^n \delta_i \cdot K(\mathbf{x}, \mathbf{x}^i)$ where $\boldsymbol{\delta} = [\delta_1, \ldots, \delta_n]^\top = (G + \lambda \cdot I)^{-1}\mathbf{y}$ where $\mathbf{y} = [y^1, \ldots, y^n]^\top$. Show that if G is invertible and we set $\lambda = 0$, then $\gamma_i = \delta_i$ for all $i \in [n]$. This means that kernel regression and landmarking-based regression will always learn the same predictor!

Problem 3 (Short Answer: 6 x 6 = 36 marks). For each of the problems, give your answer in space provided.

1. Let $\mathbf{x} = [1,\ 1]^\top, \mathbf{y} = [2,\ 1]^\top \in \mathbb{R}^2$ and let $f : \mathbb{R}^2 \to \mathbb{R}^2$ with $f(\mathbf{z}) = z_1 \cdot \mathbf{x} + z_2 \cdot \mathbf{y}$ for any $\mathbf{z} = [z_1, z_2]^\top \in \mathbb{R}^2$. Further, $\mathbf{z} = g(r) = [r^2,\ r^3]$ where $r \in \mathbb{R}$. Show how chain rule is applied here giving major steps of the calculation, write down the expression for $\frac{df}{dr}$, and also evaluate $\frac{df}{dr}$ at $r = 2$.

Name:

Roll No.: **Dept.:**

IIT Kanpur
CS771 Intro to ML
End-semester Examination
Date: November 17, 2017

2. Give an example of a Mercer kernel $K : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ and three vectors $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^2$ such that $K(\mathbf{x}, \mathbf{y}) < K(\mathbf{x}, \mathbf{z})$ and $\|\phi_K(\mathbf{x}) - \phi_K(\mathbf{y})\|_{\mathcal{H}} < \|\phi_K(\mathbf{x}) - \phi_K(\mathbf{z})\|_{\mathcal{H}}$, where $\phi_K : \mathbb{R}^2 \to \mathcal{H}$ is the feature map for the kernel K. This means that the kernel thinks $\mathbf{x}$ and $\mathbf{y}$ are less similar than $\mathbf{x}$ and $\mathbf{z}$ but in the RKHS, $\mathbf{x}$ and $\mathbf{y}$ are closer than $\mathbf{x}$ and $\mathbf{z}$. You need to give the explicit form of the kernel, the three vectors, as well as values of $K(\mathbf{x}, \mathbf{y}), K(\mathbf{x}, \mathbf{z}), \|\phi_K(\mathbf{x}) - \phi_K(\mathbf{y})\|_{\mathcal{H}}, \|\phi_K(\mathbf{x}) - \phi_K(\mathbf{z})\|_{\mathcal{H}}$ for your construction.

3. Suppose $\phi : \mathbb{R}^2 \mapsto \mathbb{R}^4$ is a linear map i.e. $\phi(\mathbf{x}+\mathbf{y}) = \phi(\mathbf{x})+\phi(\mathbf{y})$ and $\phi(c\cdot\mathbf{x}) = c\cdot\phi(\mathbf{x})$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2, c \in \mathbb{R}$. Suppose $\phi([1,\ 1]) = [1,\ 1,\ 2,\ 1], \phi([1,\ 2]) = [1,\ 2,\ 3,\ 2]$, and $\phi([2,\ 0]) = [2,\ 0,\ 2,\ 0]$. Find the matrix $M \in \mathbb{R}^{4\times 2}$ such that $\phi(\mathbf{x}) = M\mathbf{x}$ for all $\mathbf{x} \in \mathbb{R}^2$. Suppose I learn a model $\mathbf{W} = [2,\ 3,\ 1,\ 1] \in \mathbb{R}^4$. Find a model $\mathbf{w} \in \mathbb{R}^2$ such that $\langle \mathbf{w}, \mathbf{x} \rangle = \langle \mathbf{W}, \phi(\mathbf{x}) \rangle$ for all $\mathbf{x} \in \mathbb{R}^2$. Fill entries of M and $\mathbf{w}$ below.

$$M = \begin{bmatrix} \square & \square \\ \square & \square \\ \square & \square \\ \square & \square \end{bmatrix} \qquad \mathbf{w} = \begin{bmatrix} \square & \square \end{bmatrix}$$

4. Consider the following ridge regression problem $\min_{\mathbf{w}\in\mathbb{R}^d} 0.5 \cdot \|\mathbf{w}\|_2^2 + 0.5 \cdot \sum_{i=1}^n (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2$. Denote $X = [\mathbf{x}^1, \ldots, \mathbf{x}^n] \in \mathbb{R}^{d\times n}, \mathbf{y} = [y^1, \ldots, y^n]^\top \in \mathbb{R}^n$. Write down the gradient and the Hessian of the objective function at an arbitrary point $\mathbf{w} \in \mathbb{R}^d$. Then start at $\mathbf{w}^0 = \mathbf{0}$ and execute the Newton method on this problem for 3 iterations. Write down expressions for the iterates $\mathbf{w}^1, \mathbf{w}^2, \mathbf{w}^3$ that you obtain.

Name:

Roll No.: **Dept.**:

5. Recall the ϵ-insensitive loss defined as $\ell_\epsilon(y, \hat{y}) = 0$ if $|y - \hat{y}| \leq \epsilon$ and otherwise $\ell_\epsilon(y, \hat{y}) = (|y - \hat{y}| - \epsilon)^2$ where $\hat{y}, y \in \mathbb{R}$. Consider the following optimization problem with $\mathbf{x}^i \in \mathbb{R}^d, y^i \in \mathbb{R}$ and write down a likelihood distribution for $\mathbb{P}[y^i \,|\, \mathbf{x}^i, \mathbf{w}]$ and prior $\mathbb{P}[\mathbf{w}]$ such that $\hat{\mathbf{w}}$ is the MAP estimate for your model. Give explicit forms for the density functions but you need not calculate normalization constants.

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \sum_{i=1}^{n} \ell_\epsilon(y^i, \langle \mathbf{w}, \mathbf{x}^i \rangle) + \|\mathbf{w}\|_2^2$$

6. The perceptron algorithm makes the update $\mathbf{w}^{t+1} \leftarrow \mathbf{w}^t + \eta_t y^t \cdot \mathbf{x}^t$ when it misclassifies the t-th data point $(\mathbf{x}^t, y^t) \in \mathbb{R}^d \times \{-1, +1\}$. Show that if we decide to use a constant step length i.e. $\eta_t \equiv \eta$ for all t, it does not matter which value of η we choose so long as we choose a value $\eta > 0$. Specifically, show that the perceptron algorithm makes the same set of mistakes when using the constant step length η, for all $\eta > 0$.

Name:

Roll No.: **Dept.:**

IIT Kanpur
CS771 Intro to ML
End-semester Examination
Date: November 17, 2017

Problem 4 (Long Answer: $12 + 6 + 6 = 24$ marks)**.** Consider the problem of heteroscedastic regression, a curious variant of linear regression where the noise added to each data point comes from a different distribution! Let $\mathbf{x}^i \in \mathbb{R}^d, i = 1, \ldots, n$ denote the covariates/feature vectors. The responses are generated as $y^i = \langle \mathbf{w}, \mathbf{x}^i \rangle + \epsilon^i$, where the noise $\epsilon^i \sim \mathcal{N}(0, \sigma_i^2)$ for the i-th data point has variance σ_i^2 . We are shown $\{(\mathbf{x}^i, y^i)\}_{i \in [n]}$ but model $\{\sigma_i\}_{i \in [n]}$ as latent variables. Note that this is a discriminative model and $\mathbf{x}^i$ are not probabilistically modelled. You may find the shorthands $X = [\mathbf{x}^1, \ldots, \mathbf{x}^n], \mathbf{y} = [y^1, \ldots, y^n], \Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ to be helpful. Also, in all questions below, your expressions may have unspecified normalization constants. Give brief/concise derivations.

1. Derive an expression for $\mathbb{P}[\sigma_i \mid y^i, \mathbf{x}^i, \mathbf{w}]$ using the prior $\mathbb{P}[\sigma_i] = 1$ if $\sigma_i \in [0, 1]$ and $\mathbb{P}[\sigma_i] = 0$ otherwise. Then derive the MAP estimate for σ_i i.e. $\arg\max\ \mathbb{P}[\sigma_i \mid y^i, \mathbf{x}^i, \mathbf{w}]$ assuming the model $\mathbf{w}$ is known. For simplicity, assume $\mathbb{P}[\sigma_i \mid \mathbf{x}^i, \mathbf{w}] = \mathbb{P}[\sigma_i]$ i.e. $\mathbf{w}$ and $\mathbf{x}^i$ had nothing to do with the selection of σ_i.

2. Derive an expression for $\mathbb{P}[\mathbf{w} \mid y^i, \mathbf{x}^i, \sigma_i]$ using a standard Gaussian prior $\mathbb{P}[\mathbf{w}] = \frac{1}{\sqrt{(2\pi)^d}} \exp(-\frac{1}{2} \|\mathbf{w}\|_2^2)$. Then derive the MAP estimate for $\mathbf{w}$ i.e. $\arg\max\ \mathbb{P}[\mathbf{w} \mid \mathbf{y}, X, \Sigma]$ assuming that $\{\sigma_i\}$ are known.

3. Using the above estimates, give the pseudocode for an alternating optimization algorithm for estimating $\mathbf{w}$ that performs MAP-based hard assignments to the latent variables σ_i to solve the problem. Give precise update expressions in your pseudocode and not just vague statements.

Name:

Roll No.: **Dept.**:

Problem 5 (Long Answer: 8 + 16 = 24 marks). For each of the problems, give your answer in space provided.

1. Let $R \in \mathbb{R}^{d \times d}$ be a symmetric, invertible matrix, $\mathbf{x}^i \in \mathbb{R}^d$, and $y^i \in \mathbb{R}$ for $i = 1, \ldots, n$. Using the same trick we used in class of introducing a new variable $\mathbf{r}_i = y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle$ and corresponding constraints, solve the problem given below. Give 1) the Lagrangian, 2) the simplified dual optimization problem (with primal variables eliminated completely), 3) the dual solution and 4) the final primal solution $\hat{\mathbf{w}}$. Some shorthands you may find useful are $X = [\mathbf{x}^1, \ldots, \mathbf{x}^n] \in \mathbb{R}^{d \times n}$ and $H = X^\top R^{-1} X \in \mathbb{R}^{n \times n}$ i.e. $H_{ij} = (\mathbf{x}^i)^\top R^{-1} \mathbf{x}^j$.
$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{2} \mathbf{w}^\top R \mathbf{w} + \frac{1}{2} \sum_{i=1}^{n} (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2$$

2. Flopkart.com has a customer who uses his account to make purchases for his entire family. There are k members in the family, each indexed by a vector $\mathbf{u}^1, \ldots, \mathbf{u}^k \in \mathbb{R}^d$. Each product on Flopkart.com is also indexed by a vector $\mathbf{v} \in \mathbb{R}^d$. It is known that the i-th member will give the product $\mathbf{v}$, a rating $r = \langle \mathbf{u}^i, \mathbf{v} \rangle + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 1)$. The customer has made n purchases with Flopkart. In the t-th purchase, the item $\mathbf{v}^t$ was purchased and a rating r^t was given to it but it is not known which member gave that rating. We have $\{(\mathbf{v}^t, r^t)\}_{t \in [n]}$ with us. Design an algorithm to estimate the user vectors corresponding to the k members of the family. Clearly specify what are the observed and latent variables in your model and give major steps of derivation whenever your algorithm uses a MAP/MLE/other estimate. Give pseudo code of your algorithm. Avoid very fine and unnecessary details e.g. application of first order optimality.

Name:

Roll No.: **Dept.**: