CS5590: Final Exam

30-Nov-2019 (9:30am-12:30pm)

ROLL NO.________________

Important Instructions

1. Fill the blanks such that the respective statements become **true**.
2. While filling the blanks/boxes strictly follow the **instructions** in the respective question appearing immediately after/before the blank/box.
3. In case a blank is followed by a box for writing a justification for your answer, then marks will be awarded for the blank **if and only if** the answer in the blank and the justification in the corresponding box are precise.
4. Please attempt the problems in **rough sheets** first and prepare answers for all the blanks/boxes in rough. Then fair copy the required portions into this sheet while respecting the boundaries of the blanks/boxes.
5. The evaluator shall be more **strictly** following the policy of ignoring writings outside the blanks/boxes this time!
6. Please submit this booklet to the invigilator. Separately, pin all the rough sheets and submit them.
7. The rough sheets may also be scrutinized by the evaluator. So please enter question numbers and your detailed working. In case of questions that require you to work in rough, if the evaluator finds that the blanks/boxes have precise answers but there is no entry in the corresponding rough sheets, then **no** marks will be given. The rough work can still be very informal in all other aspects. Do not forget to write your **roll no.** on every rough sheet that you use.

8. All questions carry 3.5 marks. There will be **no** partial marking.

9. This paper has a choice! Overall, attempt 20 questions across the various sections. In case you attempt more, then the first 20 answers will be considered and the rest will **not** be evaluated. Please understand that this need **NOT** be same as the "best of attempted". In case you want the evaluator to not consider an answer then explicitly **strike through** it.

1 Simple yet fresh

1. Let k_1, k_2 be two linear kernels defined over R^2. k_1 is defined by $k_1(x, y) \equiv x^\top \begin{bmatrix} 3 & 1 \\ 1 & 5 \end{bmatrix} y$ and k_2 is the default linear kernel defined by $k_2(x, y) = x^\top y$ for all $x, y \in \mathbb{R}^2$. Then, the function k defined by $k \equiv k_1 - k_2$ __________ a kernel. Fill in this blank with either "is" or "is not". Justify your answer in the box below:

2. Consider the **Ramanujan** τ function, which is an integer-valued function defined over the naturals (i.e., $\tau : \mathbb{N} \mapsto \mathbb{Z}$), defined by the identity:
$$\sum_{n \in \mathbb{N}} \tau(n) q^n = q \prod_{n \in \mathbb{N}} (1 - q^n)^{24},$$
where $q(z) \equiv e^{2\pi i z}\ \forall z \ni Im(z) > 0$. Here, i is the (purely) imaginary number defined by $i^2 = -1$ and $Im(z)$ is the imaginary part of the complex number z. For example[1], $\tau(1) = 1, \tau(2) = -24, \ldots$ and so on. Then, the function $k : \mathbb{N} \times \mathbb{N} \mapsto \mathbb{Z}$ defined by, $k(n_1, n_2) =$ the product of the Ramanujan's τ function evaluated at n_1, n_2, __________ a kernel over $\mathbb{N}$. Fill in this blank with either "is" or "is not". Justify your answer in the box below:

3. Consider a classification problem, where the input space is Euclidean. *Vipareeta Buddhi* claims that the kernelized k-NN classification with Gaussian kernel will be exactly same as the *vanilla* k-NN classification[2], where the distance used is the Euclidean distance. His claim is _________. Fill in this blank with either "true" or "false". Justify your answer in the box below:

[1] It took around 60 years for mathematicians to prove the famous Ramanujan's conjecture: for all primes p, $|\tau(p)| \leq 2p^{\frac{11}{2}}$.

[2] In other words, Vanilla k-NN is kernelized k-NN with the (default) linear kernel.

4. Consider the multiquadric family of functions defined by $k_c(x, y) \equiv \sqrt{||x-y||^2 + c^2}$, $\forall\ x, y \in \mathbb{R}^n$. Here, $c \in \mathbb{R}$ is a (hyper)parameter. The value(s) of c with which k_c is a valid kernel are ____________________.

5. *Sukshma Buddhi* and *Sthula Buddhi* both plan to deloy a particular Binary classification model that has one hyperparameter, $\gamma \in (0, 1]$. Both of them use 10-fold CV for estimating γ. However, *Sthula Buddhi* considers the range of candidate values for γ as $\{1, \frac{1}{2}, \ldots, \frac{1}{2^9}\}$, whereas *Sukshma Buddhi*, being an expert in numerical optimization, considers entire interval $(0, 1]$ and finds the "best" (upto numerical errors[3]). Let us assume both have access to the same data and consider the same CV folds. Then, smaller the difference between the CV errors with the two models, ____________ likely it is that *Sthula Buddhi*'s model will perform better, when deployed, than *Sukshma Buddhi*'s model. Fill in the blank with either "more" or "less", while giving a single sentence justification in the box below:

6. With respect to k-NN models, statement(s) ______________, among the following, is(are) false, and the remaining are(is) true:

A. k-NN models are non-parametric.

B. k-NN models achieve Bayes consistency for high enough, yet finite k, provided $m \to \infty$.

C. As $m \to \infty$, the expected (true) risk with 3-NN is strictly smaller than that with 5-NN.

D. k-NN models suffer from the curse of dimensionalty, but are computationally attractive for high-dimensional data.

E. k-NN models do not suffer from curse of dimensionalty, but are computationally in-attractive for high-dimensional data.

[3] Let's assume that the CV error happens to be a "nice" function to optimize numerically over $(0, 1]$.

Fill the above blank with one or more of “A”, “B”, “C”, ”D”, “E”.

7. *Karataka* gave the following regression dataset to *Damanaka*: $\mathcal{D} = \left\{(0.1, -10), (\pi, 20), (\sqrt{2}, 5)\right\}$, where the first entry in each pair is the input and the second is the output. Further, *Karataka* instructed *Damanaka* to train a KDE-based generative regression model using this dataset. After training, *Damanaka* claims that the predicted output with the trained model at $x = 9$ is 80. *Damanaka*'s claim ______________________________. Fill this blank with either “is definitely false” or “may be true” or “cannot be validated from the given information”, while providing justification in the box below:

8. With respect to kernelized Support Vector Clustering (SVC), statement(s) ______________, among the following, is(are) true, and the remaining are(is) false:

   A. SVC models are non-parametric.

   B. SVC models suffer from the curse of dimensionalty, but are computationally attractive for high-dimensional data.

   C. SVC models do not suffer from curse of dimensionalty, but are computationally in-attractive when no. training samples are high.

   D. The no. of clusters is NOT a hyperparameter in SVC models.

   Fill the above blank with one or more of “A”, “B”, “C”, ”D”.

9. Consider the EM algorithm for MLE of GMM initialized with components as $\mathcal{N}(-1, 2), \mathcal{N}(0, 1), \mathcal{N}(1, \frac{1}{4})$, and mixing coefficients as $\frac{1}{2}, \frac{1}{3}, \frac{1}{6}$. If the training data is $(\ln(9) - 1, \ln(16) - 1, \ln(25) - 1)$, then the mixing coefficient of the first component after the first iteration (i.e., after exactly one update to the initialized value) will be ______________. Fill this blank with a number[4].

[4] Represented either in decimal format or fractions etc. Your answer may include symbols for known numerical constants.

10. Consider dimensionality reduction to a single dimension using kernelized PCA over the data $\{(1,1),(-1,1)\}$. The kernel is the quadratic one defined by $k(x,y) = (1+x^\top y)^2$. Then, the single dimensional representation of $(1,1)$ is ______. Fill this blank with a number[5].

11. In lectures, we tried to visualize how a neural network that corresponds to logistic regression will look like in case of binary classification and regression. Now we would like to visualize a network that corresponds to 3-class Bayes classifier, where the class conditionals are modeled using different exponential family models. Please draw a visualization of such a (feed-forward neural) network, with depth 2, in the following box. Clearly mention the activation function (of your choice that is appropriate) at each hidden layer as well as the output layer. Also, write down the loss function in terms of the values (activations) at the output layer:

12. Recall that the formula[6] for the representation, $z \in \mathbb{R}^r$, for a datapoint, $x \in \mathbb{R}^n$, is given by $z = \Lambda^{-\frac{1}{2}} L^\top (x - \hat{\mu})$ in the PPCA model. Here, Λ is the diagonal matrix with entries as largest r eigenvalues of the sample covariance[7], L is the matrix whose columns are the

[5] Represented either in decimal format or fractions etc. Your answer may include symbols for known numerical constants.

[6] Here, for the sake of simplicity, I present the formula in the asymptotic case, where noise variance → 0.

[7] Let's assume these r values are positive. Also, unlike in PCA, these are not eigenvalues

corresponding (normalized) eigenvectors, and $\hat{\mu}$ is the sample mean. Is PPCA kernelizable? If yes, present a pseudocode for kernelized PPCA in the box below[8]. Else, explain why exactly PPCA can't be kernelized, while PCA can be!

13. Consider a regression problem where the loss function is defined by $l(y, y^{'}) \equiv \left(y - y^{'}\right)^4$. With a particular probabilistic model the estimated posterior likelihood, $\hat{p}(y/x)$, turned out be such that the moments $\mathbb{E}_{\hat{p}}[Y^d/x] = \max(d-2, 0)x$. Then, the corresponding estimate for the Bayes optimal at $x = 2$ would be ____. Fill this blank with a number[9].

14. Consider a supervised learning problem where the input and output spaces are both $\mathbb{R}$. Consider the discriminative model induced by the exponential family, with the sufficient statistics (over outputs) as $\psi(y) = \begin{bmatrix} y \\ y^3 \end{bmatrix}$ and the feature map (over inputs) as $\phi(x) = \begin{bmatrix} \sin(x) \\ \cos(x) \end{bmatrix} \in \mathbb{R}^2$. Also, consider a generative model for the same problem with sufficient statistics as $\theta(x, y)$. If

$$\theta(x, y) \equiv \qquad\qquad\qquad ,$$

of the sample correlation matrix, but of the sample covariance itself!

[8] You don't need to justify the correctness of your pseudocode.

[9] Represented either in decimal format or fractions etc. Your answer may include symbols for known numerical constants.

then the form of the posterior likelihood with the generative model will be exactly same as that for the likelihood modeled in the discriminative model.

[3.5Marks×14=49 Marks]

2 Recycled from Practice set

1. Consider the function $k : \{1, \ldots, n\} \times \{1, \ldots, n\} \mapsto \mathbb{R}$ defined by $k(x, z) \equiv \min(x, z), \forall\ x, z \in \{1, \ldots, n\}$. Is k a valid kernel? In either case, please provide supporting arguments in the box below:

2. Consider the function $k : \mathbb{R}^{++} \times \mathbb{R}^{++} \mapsto \mathbb{R}$ defined by $k(x, z) \equiv \frac{1}{x+z}, \forall\ x, z > 0$. Is k a valid kernel? In either case, please provide supporting arguments in the box below:

3. Let $\mathcal{G}_1, \ldots, \mathcal{G}_k$ be inductive-bias sets (hypothesis classes), each of which can be employed in a particular learning task. Consider the following two learners:

   L1: The learner that employs $\cup_{i=1}^{k} \mathcal{G}_i$ as the inductive bias and ERM for parameter estimation.

   L2: The learner that employs validation for model selection among $\mathcal{G}_1, \ldots, \mathcal{G}_k$ and ERM for parameter estimation.

   Describe a scenario in which $L1$ may generalize better than $L2$ in the box below:

4. Consider a regression problem with input space as $\mathbb{R}^n$ and a generative model that is a mixture model. In the box below, argue that the posterior is also a mixture model.

5. Consider the following stochastic optimization problem with loss function, l, as the squared-loss:

$$\min_{||w||=1} \mathbb{E}[l(X, ww^\top X)]$$

In the box below, provide a simplified (and geometrically appealing) interpretation to it's ERM solution, along with justification:

6. In the box below provide a simple example of training data where the solution obtained by k-means algorithm is not a local minimum. *Hint:* Think about 1-d data and $k = 2$. Also, provide a concise justification.

[3.5Marks×6=21 Marks]

3 Reappearing Ones :)

1. Consider a regression problem where the input space, $\mathcal{X} = \mathbb{R}^n$, and the output space, $\mathcal{Y} = \mathbb{R}$. It is proposed to use the inductive bias as the set of all affine functions:

   $$\mathcal{G} \equiv \left\{ g \mid \exists\ w \in \mathbb{R}^n,\ b \in \mathbb{R}\ \ni\ g(x) = w^\top x - b\ \forall\ x \in \mathcal{X} \right\}.$$

   The parameters are $w \in \mathbb{R}^n, b \in \mathbb{R}$. The loss to be used is the square loss. Let the input vectors in the training set be arranged as column vectors in the matrix $X_{n\times m}$, where m is the training set size. Let $y_{m\times 1}$ denote the vector with entries as the corresponding outputs in the training set. Let us denote the q-dimensional vector with all entries as unity by 1_q and the identity matrix of size n by I_n. Then, the simplified expressions for the solution, $\left(\hat{w}, \hat{b}\right)$, of the corresponding ridge-regression problem are given by:

   $\hat{w} =$ ____

   $\hat{b} =$ ______________________ .

   Your expression for $\hat{w}$ must only be in terms of $X, y, I_n, 1_m, m, \lambda$, where λ is the hyperparameter multiplying the regularizer term. And, your expression for $\hat{b}$ must involve $\hat{w}, X, y, 1_m, m$ alone.

2. In the box below write the pseudocode for the (batch) perception algorithm as mentioned in your textbook:

3. In the box below repeat the derivation of the parameter estimation method called MLE from first principles, as done in the lectures:

4. In the box below repeat the rough derivation from the lectures that shows that the (true) risk with 1-NN is bounded above by twice of that with the Bayes optimal, in the asymptotic case.

[3.5Marks×4=14 Marks]