Name:

Roll No.:

Dept.:

Instructions:

Total: **40 marks**

1. Please write your name, roll number, department on **both sides** of this question paper.

Section 1 (20 problems: 20 x 2 = 40 marks). Write your answers precisely and concisely in the provided space.

1. Consider learning a regression model with N training examples $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$. Write down a regularized loss function that is robust against outlier examples *and* also gives a sparse regression weight vector.

   $\sum_{n=1}^{N} |y_n - W^T x_n| + \lambda \|W\|_1$ (Absolute loss + L1 regularizer)

2. Suppose you've learned a linear SVM model $\boldsymbol{w} \in \mathbb{R}^D$ and $b \in \mathbb{R}$. How'd you use it to compute the *probability* of a test input $\boldsymbol{x}$'s label y being 1? Clearly write down the expression to compute this probability.

   Similar to logistic regression: $P(y=1|x,w,b) = \frac{1}{1+\exp[-(w^Tx+b)]}$

3. Consider a generative classification model for binary classification. Assume both classes to be modeled by Gaussians with equal covariances. For a point $\boldsymbol{x}_*$ *exactly* at the middle of the line joining their means, would the following be true: $p(y=1|\boldsymbol{x}_*) = p(y=-1|\boldsymbol{x}_*) = 0.5$? Briefly justify your answer.

   No, because $P(y=k|x_*) \propto P(y=k)P(x_*|y=k)$ ← Being in the middle will only make this term equal.

   π_k ← The class marginal matters too!

4. State whether the following statement is true or false: Training error of one-nearest neighbor can never be more than that of three-nearest neighbors. You also need to briefly justify your answer.

   Yes, because training error of 1-NN is zero.

5. Consider a regression loss function $\mathcal{L}(\boldsymbol{w}) = \sum_{n=1}^N \gamma_n (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2 + \sum_{d=1}^D \lambda_d w_d^2$. What would be the probability distributions for the likelihood $p(y_n|\boldsymbol{w}, \boldsymbol{x}_n)$ and the prior $p(\boldsymbol{w})$ for this model? The exact form of these distributionss are not required but clearly mention the parameters of these distributions.

   $P(y_n|w,x_n) = N(y_n|W^Tx_n, \gamma_n^{-1})$, $P(w) = N(w|0,\Lambda^{-1})$ with $\Lambda = \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \ddots \\ & & & \lambda_D \end{bmatrix}$

6. A symmetric $N \times N$ matrix $\mathbf{M}$ is positive semi-definite (p.s.d.) if $\boldsymbol{z}^\top \mathbf{M} \boldsymbol{z} \geq 0$, for all vectors $\boldsymbol{z} \in \mathbb{R}^N$. Given an $N \times D$ matrix $\mathbf{X}$, is the matrix $\mathbf{X}\mathbf{X}^\top$ p.s.d.? Prove your answer.

   $Z^TXX^TZ = (X^TZ)^T(X^TZ) = S^TS \geq 0$ where $S = X^TZ \in \mathbb{R}^D$

7. Using hard-margin SVM's Lagrangian $\mathcal{L}(\boldsymbol{w}, b, \alpha) = \frac{\boldsymbol{w}^\top \boldsymbol{w}}{2} + \sum_{n=1}^N \alpha_n \{1 - y_n(\boldsymbol{w}^\top \boldsymbol{x}_n + b)\}$, show that the sum of Lagrange multipliers of positive examples = sum of Langrange multipliers of negative examples.

   Taking deriv. w.r.t b, we get $\sum_{n=1}^{N} \alpha_n y_n = 0 \Rightarrow \sum_{n:y_n=1} \alpha_n = \sum_{n:y_n=-1} \alpha_n$

8. In a generative classification model with class-conditional distributions $p(\boldsymbol{x}|y=k) = \mathcal{N}(\boldsymbol{x}|\mu_k, \Sigma_k)$ and class-marginals $p(y=k) = \pi_k$, $k = 1, \ldots, K$, what's the marginal distribution $p(\boldsymbol{x})$ of the inputs?

   $P(x) = \sum_{k=1}^{K} P(x, y=k) = \sum_{k=1}^{K} P(y=k)P(x|y=k) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)$

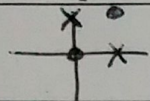
Name:

Roll No.: Dept.:

IIT Kanpur
CS771 Intro to ML
Quiz
Date: October 3, 2018

---

9. Consider 4 training examples $\{(0,0),(1,0),(0,1),(1,1)\}$ with labels $\{+1,-1,-1,+1\}$. How many iteration will Perceptron take to converge on this data? Briefly justify your answer.



Data is linearly NOT seperable. Perceptron will NEVER converge.

10. Write down the loss function for a K class prototype classification model, given N training examples $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$. The unknown parameters in the loss function are the means $\mu_1, \ldots, \mu_K$ of the K classes.

· Many ways to write it. Just like K-means, with known labels,
$\mathcal{L}(\mu_1, \ldots, \mu_K) = \sum_{n=1}^{N} \|x_n - \mu_{y_n}\|^2$ (Can also use any of the forms to K-means loss func.)

11. Would solving SVM using SGD give the same solution as Perceptron? If yes, why? If no, why not?

No, because SVM's loss function is not the same as Perceptron's loss function. (Also, SVM maximizes the margin, Perceptron doesn't)

12. Why might you want to solve linear regression using gradient descent instead of in closed form?

Because closed form solution requires expensive matrix inversion.

13. Suppose we know that the weight vector $\boldsymbol{w}$ of a linear/logistic regression model is close to a known vector $\boldsymbol{w}_0$. How would you use this information (1) As a regularizer, (2) As a prior distribution?

Regularizer: $\|w - w_0\|^2$ or $\|w - w_0\|_1$
Prior : $\mathcal{N}(w \mid w_0, \lambda^{-1} I)$ or Laplace prior with mean w_0.

14. Consider the landmark based approach for getting explicit features from a kernel k. Given N training inputs $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, what will be the landmark based feature vector if each input is a landmark point?

$\psi(x_n) = [k(x_n, x_1) \ldots k(x_n, x_N)] \in \mathbb{R}^N$

15. Consider a regularized model with loss function $\sum_{n=1}^N \ell(y_n, \boldsymbol{w}^\top \boldsymbol{x}_n) + \lambda\|\boldsymbol{w}\|^2$. What happens to the training error when λ is set to (1) a very very small value, and (2) a very very large value?

(1) Training error will become very very small (overfit on training data)
(2) Training error will become very very large (Too much regularization) UNDERFITTING

16. Rank gradient descent (GD), stochastic gradient descent (SGD), and mini-batch gradient descent (MGD) in terms of per-iteration cost. Briefly justify your ranking.

(1) SGD (2) MGD (3) GD
FASTEST SLOWEST
one possibility

17. A possible vector $\boldsymbol{w} \in \mathbb{R}^4$ with $\|\boldsymbol{w}\|_1 = 0.5$, $\|\boldsymbol{w}\|_0 = 2$, and $\sum_{d=1}^4 w_d = 0$, will be $\boldsymbol{w} =$ $[0.25, 0, -0.25, 0]^\top$

18. SGD update for ridge regression $\sum_{n=1}^N (y_n - \boldsymbol{w}^\top \boldsymbol{x}_n)^2 + \lambda\|\boldsymbol{w}\|^2$: $w^{(t+1)} = w^{(t)} + 2\eta_t [(y_n - w^{(t)\top} x_n) x_n + \lambda w^{(t)}]$

19. Mark all options (by encircling them in bold) that are true: (1) Prototype classification can be kernelized, (2) Kernelized SVM is slower than kernelized Perceptron at test time, (3) Kernel K-means is slower than standard K-means, (4) Training SVM with RBF kernel is more expensive than with quadratic kernel.

20. Mark all options (by encircling them in bold) that are true about the logistic regression model which uses $p(y=1|\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{1+\exp(\boldsymbol{w}^\top \boldsymbol{x})}$: (1) Can't solve for $\boldsymbol{w}$ in closed form, (2) Can't be kernelized; (3) GD will give the same solution regardless of initialization, (4) Gaussian prior on $\boldsymbol{w}$ makes deriving the posterior easy.