CS771 Quiz 1

BHAVY KHATRI

TOTAL POINTS

29.5 / 40

QUESTION 1

1 Q1 0 / 2

+ **1 pts** Absolute loss function
+ **1 pts** L1 or L0 regularizer (or L_p with p < 1) on weight vector
✓ + **0 pts** **Not attempted or incorrect.**

QUESTION 2

2 Q2 2 / 2

✓ + **2 pts** **Transforms the score w'x+b into a probability using some function that gives a number between 0 and 1. The function should be such that very large negative score should lead to prob. close to 0 and very large positive score should lead to prob. close to 1.**
+ **1 pts** Basic idea correct but some minor errors (e.g. the signs are incorrect or if b is missing from the expression).
+ **0 pts** Not attempted or incorrect.

QUESTION 3

3 Q3 0 / 2

+ **2 pts** Answers that it will be FALSE and gives the justification that the class marginals also matter.
+ **2 pts** Mentions True if class marginals are equal.
+ **0 pts** Answers that it will be FALSE but the justification is wrong (does not mention class marginals or anything related to it)
✓ + **0 pts** **Incorrect.**
+ **0 pts** Not attempted.

QUESTION 4

4 Q4 0 / 2

+ **2 pts** Answers as TRUE and gives the correct justification (i.e., the training error of 1-NN is zero)
+ **1 pts** Answers as TRUE but justification not entirely proper.
✓ + **0 pts** **Incorrect.**
+ **0 pts** Not attempted.
+ **0.5 pts** Only mentions True without justification
+ **0.5 pts** Mentions True with wrong or vague justification.

QUESTION 5

5 Q5 1 / 2

+ **1 pts** Correct likelihood (univariate Gaussian with mean $$w^Tx_n$$ and precision $$\gamma_n$$)
+ **1 pts** Correct prior (multivariate, zero mean Gaussian with diagonal covariance with d-th diagonal entry = $$\lambda_d^{-1}$$)
✓ + **0.5 pts** **If likelihood expression is mostly correct but some errors ({e.g., $$\gamma_n$$ used as variance, not precision)**
✓ + **0.5 pts** **If prior is mostly correct but some errors**
+ **0 pts** Not attempted or incorrect.

QUESTION 6

6 Q6 2 / 2

✓ + **2 pts** **Answers YES and gives the correct proof.**
+ **1 pts** Answers YES but has minor errors in the proof.
+ **0 pts** Not attempted or incorrect.

QUESTION 7

7 Q7 2 / 2

✓ + **2 pts** **Uses (or derives) \sum_{n=1}^N \alpha_n y_n = 0$ and correctly shows the results by separating terms with positive and negative y_n.**
+ **1 pts** Seems to use the basic idea correctly but proof is not properly done (it's a very short proof anyway)

+ **0 pts** Not attempted or incorrect.

QUESTION 8

8 Q8 2 / 2

✓ + **2 pts** **Writes the correct expression directly or shows how to get it and then writes the expression**

+ **1 pts** Uses the correct idea (marginalization) but not properly done.

+ **0 pts** Not attempted or incorrect.

QUESTION 9

9 Q9 2 / 2

✓ + **2 pts** **Says that it would never converge and gives the correct reason (data is not linearly separable)**

+ **0 pts** Not attempted or incorrect answer.

QUESTION 10

10 Q10 2 / 2

✓ + **2 pts** **Correct expression for the loss function (basically any of the various forms of the K-means loss function but with z_n = y_n)**

+ **0 pts** Not attempted or incorrect.

QUESTION 11

11 Q11 2 / 2

✓ + **2 pts** **Answers NO and gives correct justification (e.g., they have different objectives/loss function, or SVM maximizes the margin whereas Perceptron doesn't)**

+ **0 pts** Not attempted or incorrect.

+ **0.5 pts** Perceptron expression is correct

+ **0.5 pts** SVM expression is correct

+ **0 pts** wrong answer

+ **1.5 pts** Answer is correct but reasoning is partial correct

+ **1 pts** reasoning is correct but conclusion is wrong

+ **1 pts** partial correct but reasoning is wrong

+ **0.5 pts** only answered as no

QUESTION 12

12 Q12 2 / 2

✓ + **2 pts** **Mentions that closed form solution is computationally expensive due to matrix inversion where GD doesn't require any inversion.**

+ **0 pts** Not attempted or incorrect.

+ **1 pts** partial marks

QUESTION 13

13 Q13 2 / 2

✓ + **1 pts** **Correct regularizer**

✓ + **1 pts** **Correct prior (any distribution e.g., Gaussian or Laplace with mean = w_0)**

+ **0 pts** Not attempted or incorrect.

QUESTION 14

14 Q14 2 / 2

✓ + **2 pts** **Correct expression**

+ **0 pts** Not attempted or incorrect expression

+ **1 pts** Partially correct expression (e.g. dot pdt. instead of kernel function)

QUESTION 15

15 Q15 1 / 2

+ **2 pts** Correctly mentions that for case (1) the training error will become very small and for case (2) the training error will become very large.

+ **1 pts** Only part (1) correct, i.e., training error will become very very small

+ **1 pts** Only part (2) correct, i.e., training error will become very very large

✓ + **1 pts** **Mentions overfitting for part (1) and underfitting for part (2) but doesn't say what it implies for training error (what the question asked).**

+ **0 pts** Not attempted or incorrect answer.

QUESTION 16

16 Q16 2 / 2

✓ + **2 pts** **Correct ordering and proper justification (GD uses all the data, MGD uses a minibatch, SGD uses a single example)**

- **0.5 pts** Justification has some minor errors

+ **0 pts** Not attempted or incorrect.

QUESTION 17

17 Q17 2 / 2

✓ **+ 2 pts** **Correct solution (multiple solutions are possible but it should have two 0s, one 0.25 and one -0.25).**

+ 1 pts Incorrect solution, but L1 and L0 norms are satisfied.

+ 0 pts Not attempted or incorrect.

QUESTION 18

18 Q18 2 / 2

✓ **+ 2 pts** **Correct expresion**

+ 1.5 pts Mostly correct expression but sign is flipped

+ 1 pts Missing the gradient of the regularizer but otherwise correct.

+ 1 pts Incomplete/partially correct expression

+ 0 pts Not attempted or incorrect.

QUESTION 19

19 Q19 1 / 2

+ 2 pts If both correct (and only those) options are encircled.

+ 1 pts If only one of the correct options is encircled, and no other option correct/incorrect is encircled.

✓ **+ 1 pts** **If both correct options are encircled but an incorrect option is also encircled**

+ 0 pts All options encircled or all options skipped.

+ 0 pts Not attempted or incorrect.

QUESTION 20

20 Q20 0.5 / 2

+ 2 pts If both correct (and only those) options are encircled.

+ 1 pts If only one of the correct options is encircled, and no other option correct/incorrect is encircled.

+ 1 pts If both correct options are encircled but an incorrect option is also encircled

+ 0 pts All options encircled or all options skipped.

+ 0 pts Not attempted or incorrect.

✓ **+ 0.5 pts** **One correct and one incorrect**

Name: BHAVY KHATRI

Roll No.: 150186 Dept.: MTH

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

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.

$\mathcal{L}(\omega) = \sum_{n=1}^{N} (y_n - \omega^T x_n)^2 + \frac{\lambda}{2}\omega^T\omega, \quad \omega = (X^T X + \lambda I_D)^{-1} X^T y$

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.

$P(y=1) = \frac{\exp(\omega^T x + b)}{1 + \exp(\omega^T x + b)}$

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.

let $P(y=1|x_*) = \pi$, $p(y=-1|x_*) = 1-\pi$, $p(x_*|y=1) = N(\mu_1, \Sigma)$ $p(x_*|y=-1) = N(\mu_2, \Sigma)$

$x_* = \frac{\mu_1 + \mu_2}{2}$, [True, as $\pi = 1 - \pi \Rightarrow \pi = 0.5$]

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.

False, as there is no training involved in k-nearest neighbours. We directly test the input by considering all training example, with majority of k nearest class.

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.

$\cancel{p(\omega) = \prod_{d=1}^{D} \exp(w_d, \lambda_d)}$ $\cancel{p(y_n|\omega, x_n) = \exp(y_n, \gamma_n)}$

$P(y_n|\omega, x_n) = \mathcal{N}(\omega^T x_n, \frac{1}{\sqrt{\gamma_n}})$ $\cancel{\exp(-2\gamma_n y_n}$ $P(\omega) = \prod_{d=1}^{D} \mathcal{N}(0, \frac{1}{\sqrt{\lambda_d}})$

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.

Yes XX^T is psd. as $z^T X X^T z = (X^T z)^T (X^T z) = \omega^T \omega \geq 0$

let $X^T z = \omega$ [$X^T z = D \times 1$ matrix].

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.

$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum \alpha_n y_n = 0$ $\quad \alpha_n^+ = \{\alpha_n : y_n = 1\}$ $\alpha_n^- = \{\alpha_n : y_n = -1\}$.

$\Rightarrow \boxed{\sum_{y_n = 1} \alpha_n^+ = \sum_{y_n = -1} \alpha_n^-}$

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?

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

Name: BHAVY KHATRI

Roll No.: 150186 Dept.: MTH

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.

Perceptron algo will never converge as the data is not linearly separable. Perceptron will converge iff data is linearly separable.

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.

$\cancel{\mathcal{L} = \sum_{n=1}^{N} \|x_n}$ $\mathcal{L} = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \|x_n - \mu_k\|^2$ $\quad z_{nk} = \begin{cases} 1 & \text{if } y_n = k \\ 0 & \text{o.w.} \end{cases}$

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

No, SVM using SGD gives the hyperplane with maxm. margin around two classes while ~~SGD~~ Perceptron can give any hyperplane based on the initialization of weight vector.

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

- inverse is hard to compute ~~to~~ i.e. $O(D^3)$
- multiplication of $X^T X$ takes $O(ND^2)$ time

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?

(1) Regularise $= \lambda \|w - w_0\|^2$, and $L(w) = \sum l_n(w) + \lambda \|w - w_0\|^2$

(2) $p(w) = \mathcal{N}(w_0, \sigma^2)$ $\sigma^2 \approx 0$, assume normal distribution with mean w_0 and very small variance

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?

~~(1) Regularise $= \lambda \|w - w_0\|^2$, and $L(w) = \sum l_n(w) + \lambda \|w - w_0\|^2$~~

~~(2) $p(w) \sim \mathcal{N}(w_0, \sigma^2)$ $\sigma^2 \approx 0$, assume normal distribution with mean w_0 and very small variance~~

$\phi(x_k) = [k(x_k, x_1) \ldots k(x_k, x_N)]^T \quad \forall k = 1, \ldots N$

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) Overfitting

(2) 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.

GD > MGD > SGD, GD - takes all examples into account for each iteration
MBGD - " some " " " " "
SGD - " one " " " " "

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 $\boldsymbol{w} =$ $(0.25, -0.25, 0, 0)$

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

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.