CS771 Mid-semester Examination

RAKTIM MITRA

TOTAL POINTS

78.5 / 80

QUESTION 1

1 True or False 8 / 8

+ 0 Correct
+ **8 Point adjustment**

QUESTION 2

Ultra Short Answer 24 pts

2.1 Problem 2.1 4 / 4

+ **4 Correct**
+ 0 Incorrect

2.2 Problem 2.2 4 / 4

+ **4 Correct**
+ 0 incorrect
+ 0 Not attempted

2.3 Problem 2.3 4 / 4

+ **4 Correct**
+ 0 incorrect
+ 3 d missing from complextity term
+ 2 only one expression correct

2.4 Problem 2.4 4 / 4

+ 1 part1 exp
+ 1 part ans
+ 1 part2 exp
+ 1 part2 ans
+ **4 Correct**
+ 0 incorrect

2.5 Problem 2.5 4 / 4

+ **1 Root**
+ **1 Squared error**
+ **2 Averaged**
+ 1 Absolute deviation
+ 0 Incorrect

2.6 Problem 2.6 3 / 4

+ **2 Repeated assignment of clusters**
+ 1 Mention trapped at local optima
+ 1 Complete answer
+ **1 Not fully formed arguments**
+ 0 Incorrect or wrong logic

QUESTION 3

Short Answer 32 pts

3.1 Problem 3.1 8 / 8

+ **4 Correct line in the plot**
+ **4 Correct Expression**
+ 0 Wrong Solution
+ 0 No solution
+ 2 Incomplete expression

3.2 Problem 3.2 8 / 8

+ **8 Correct**
+ 1 Some Condition(s) mentioned, but doesn't solve the problem correctly.
+ 1 Only condition(s) mentioned, without a hint of how that would solve the problem
+ 7 Correct approach, but use of < instead of <= as required for the definition of the set.
+ 7 Correct Approach
- 1 Convexity condition incorrect
+ 4 Condition(s) mentioned correct to some extent. Solution incomplete/missing details
+ 0 Not attempted/Doesn't count/Doesn't make sense

3.3 Problem 3.3 8 / 8

+ **3 Correct likelihood expression**
+ **3 correct prior expression for case ||W|| <= 1**
+ **2 correct prior expression for case ||W|| > 1**
+ 1 Partial answer for prior expression
+ 0 Wrong Answer or Irrelevant Answer or No Answer
- 1 Silly Mistakes, Neglecting Constants(like normalization), not defining prior expression

(distribution) properly, neglecting variance term..

+ **2** Partially correct answer for likelihood expression

3.4 Problem 3.4 8 / 8

+ **8 Correct**
- **2** Incorrect Cluster allocation(Should be sigma inverse)
+ **0** Not Attempted
- **3** No Algorithm.
+ **0** In correct
- **1** No need to update co-variance matrix
+ **6** No mean updation/ incorrect mean updation.
+ **3** No need to update the co-variance matrices.
+ **2** Incorrect
+ **4** Use Gaussian Model as Probability Model.

QUESTION 4

Long Answer 16 pts

4.1 Problem 4.1 3 / 3

+ **2 Argmin expression**
+ **0.5 Positivity constraints**
+ **0.5 Sum of probabilities equals to 1**
+ **0** not attempted
+ **0** wrong expression

4.2 Problem 4.2 2.5 / 3

+ **2 log probabilities**
+ **0.5** Positivity constraint
+ **0.5 Sum of probabilities constraint**
- **0.5** Sign mistake
+ **0** wrong expression
+ **0** not attempted

4.3 Problem 4.3 5 / 5

+ **0** Unattempted or wrong answer
+ **1 Correct use of Lagrangian term in dual**
+ **1 Correct order of max and min operations**
+ **2 Steps to show elimination of primal variable**
+ **1 Correct dual with primal variable eliminated**

4.4 Problem 4.4 5 / 5

+ **0** Incorrect or unattempted
+ **4 Correct expression for dual variable(s)**
+ **1 Correct expression for MLE estimate for \pi_k**

Name: Raktim Mitra

Roll No.: 150562 Dept.: CSE

IIT Kanpur
CS771 Intro to ML
Mid-semester Examination
Date: September 21, 2017

Instructions: *Total:* **80 marks**

1. This question paper contains a total of 6 pages (6 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: 8 X 1 = 8 marks). For each of the following simply write **T** or **F** in the box.

1. F — The Bayesian predictive posterior has a nice closed form solution if we have a logistic likelihood for $\mathbb{P}[y \mid \mathbf{x}, \mathbf{w}]$ and a Gaussian prior for $\mathbb{P}[\mathbf{w}]$.
2. F — Hard assignment alternating optimization approaches are much more expensive to execute than soft assignment alternating optimization approaches.
3. F — In ridge regression $(\arg\min \lambda/2 \cdot \|\mathbf{w}\|_2^2 + \|X^\top \mathbf{w} - \mathbf{y}\|_2^2)$, no matter how large a regularization constant $\lambda > 0$ we set, we will always get good solutions.
4. T — When deriving MLE solutions, working with log-likelihood terms is simpler than working with likelihood terms directly.
5. F — It is okay to perform minor evaluations on the test set during training so long as we don't do it too many times.
6. F — If S_1 and S_2 are two convex sets in $\mathbb{R}^2$, then their union $S_1 \cup S_2$ is always a convex set as well.
7. F — It is not possible to execute the SGD algorithm if the objective function is not differentiable.
8. T — Convex optimization problems like ridge regression are not as sensitive to proper initialization (while carrying out optimization) as are non-convex problems like k-means.

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

1. Suppose I have a coin with bias p i.e. it lands heads with probability p. What is the probability that when this coin is tossed n times, we observe x heads and $n - x$ tails? Give only the final expression.

$^{n}C_{x}\, p^{x}(1-p)^{n-x}$

2. Given a vector $\mathbf{a} \in \mathbb{R}^d$, what is the trace of the matrix $A = \mathbf{a}\mathbf{a}^\top \in \mathbb{R}^{d \times d}$?

The trace is l_2 norm of a squared. i.e. $\sum_{i=1}^{d} a_i^2 = \|a\|_2^2$

Name: Raktim Mitra

Roll No.: 150562 Dept.: CSE

IIT Kanpur
CS771 Intro to ML
Mid-semester Examination
Date: September 21, 2017

3. Give the time complexity of predicting the label of a new point using the OvA and AvA approaches in a multiclassification problem with K classes with d-dimensional features. Briefly justify your answer.

OvA: $\hat{W}$ is $K \times d$, we multiply it with $d \times 1$ new point and take the maximum. $\Rightarrow O(Kd)$

AVA we have $K \times K$ d-dimensional weight vectors here. Multiplying all these to $d \times 1$ new point and getting required value makes it $O(K^2 d)$

4. We are given that $\mathbb{P}[\Theta] = 0.1, \mathbb{P}[y \mid \mathbf{x}, \Theta] = 0.4, \mathbb{P}[\mathbf{x} \mid y, \Theta] = 0.5, \mathbb{P}[y \mid \Theta] = 0.2, \mathbb{P}[\mathbf{x}, y] = 0.5$. Find $\mathbb{P}[\Theta \mid \mathbf{x}, y]$ and $\mathbb{P}[\mathbf{x} \mid \Theta]$. Show your expressions for these terms briefly and the final answer.

$$P[\theta \mid x, y] = \frac{P[X \mid y, \theta]\, P[y \mid \theta]\, P[\theta]}{P[x, y]} = \frac{0.5 \times 0.2 \times 0.1}{0.5} = 0.02 \text{ [Answer]}$$

$$P[X, \theta] = \frac{P[X \mid y, \theta]\, P[y \mid \theta]\, P[\theta]}{P[Y \mid X, \theta]} \Rightarrow P[X \mid \theta] = \frac{P[X, \theta]}{P[\theta]}$$

$$= \frac{P[X \mid y, \theta]\, P[y \mid \theta]\, \cancel{P[\theta]}}{P[Y \mid X, \theta]\, \cancel{P[\theta]}}$$

$$= \frac{0.5 \cdot 0.2}{0.4} = 0.25 \text{ [Answer]}$$

5. Consider a regression problem with covariates $\mathbf{x}^i \in \mathbb{R}^d$ and responses $y^i \sim \mathcal{N}(\langle \mathbf{w}, \mathbf{x}^i \rangle, \sigma^2)$. Suppose you are given $(\mathbf{x}^i, y^i)_{i=1,2,\ldots,n}$ as well as $\mathbf{w}$. Write down an estimator for σ.

σ is ~~squared~~ average of errors in estimation squared.

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y^i - Wx^i)^2}$$

6. Let $\mathbf{z}^t \in [K]^n$ denote the cluster assignments made by the k-means algorithm at the t-th iteration i.e. data point $i \in [n]$ gets assigned to the cluster $\mathbf{z}_i^t \in [K]$. Suppose we have $\mathbf{z}^t \neq \mathbf{z}^{t+1}$ but $\mathbf{z}^t = \mathbf{z}^{t'}$ for some $t' > t + 1$? What must be happening if cluster assignments get repeated in this manner?

That means, given the initialization and point set, it is not possible to converge to a cluster set.

The function is oscillating b/w ~~some opti~~ more than one optimal solutions, but no solution leads to convergence.

Name: Rakttim Mitra

Roll No.: 150562

Dept.: CSE

IIT Kanpur
CS771 Intro to ML
Mid-semester Examination
Date: September 21, 2017

Problem 3 (Short Answer: 4 x 8 = 32 marks). For each of the problems, give your answer in space provided.

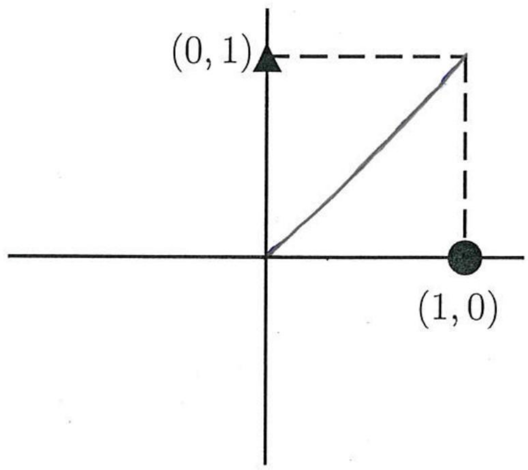
1. We wish to perform binary classification when we have two prototypes: the triangle prototype $(0,1)$ and the circle prototype $(1,0)$. Find the decision boundary when we use the L_1 metric to calculate distances i.e. $d(\mathbf{z}^1, \mathbf{z}^2) = \|\mathbf{z}^1 - \mathbf{z}^2\|_1 = |\mathbf{z}_1^1 - \mathbf{z}_1^2| + |\mathbf{z}_2^1 - \mathbf{z}_2^2|$ for $\mathbf{z}^1, \mathbf{z}^2 \in \mathbb{R}^2$. Calculate the decision boundary only within the box $B := \{\mathbf{z} \in \mathbb{R}^2 : \mathbf{z}_1, \mathbf{z}_2 \in [0,1]\} \subset \mathbb{R}^2$ and write its expression below. Draw the decision boundary in the figure. Note that you dont have to calculate the decision boundary outside the box B.

$|x| + |1-y| = |1-x| + |y|$

in box B, $x + 1 - y = 1 - x + y \Rightarrow 2x = 2y$

$\Rightarrow y = x$

or, $z_1 = z_2$ in z notation.



2. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable convex function on $\mathbb{R}^d$. Prove (with detailed steps) that the set $S_f := \{\mathbf{x} : f(\mathbf{x}) \leq 0\}$ is always a convex set. Use any definition of convexity you are comfortable with.

By convexity, $f(\alpha u + (1-\alpha)v) \leq \alpha f(u) + (1-\alpha) f(v) \quad \forall u, v \in \mathbb{R}^d \ \forall \alpha \in [0,1]$

<u>Claim</u>: $S_f : \{x : f(x) \leq 0\}$ is a convex set.

i.e. $\forall u, v \in S_f, \forall \alpha \in [0,1]$

$\alpha u + (1-\alpha) v \in S_f$ i.e $f(\alpha u + (1-\alpha)v) \leq 0$.

<u>Proof</u>: $f(\alpha u + (1-\alpha)v) \leq \alpha f(u) + (1-\alpha) f(v)$

now, $f(u) \leq 0$, $f(v) \leq 0 \quad \forall u, v \in S_f$.

$\alpha \in [0,1] \Rightarrow \alpha \geq 0$ always and $\alpha \leq 1$ always.

$\Rightarrow (1-\alpha) \geq 0$

$\Rightarrow \alpha f(u) + (1-\alpha) f(v) \leq 0$ always.

$\Rightarrow f(\alpha u + (1-\alpha)v) \leq 0$ ~~always~~

$\Rightarrow \alpha u + (1-\alpha) v \in S_f$. $\forall \alpha \in [0,1]$, $u, v \in S_f$.

$\Rightarrow S_f$ is always a convex set. [Proved]

Name: Raktim Mitra

Roll No.: 150562

Dept.: CSE

IIT Kanpur
CS771 Intro to ML
Mid-semester Examination
Date: September 21, 2017

3. Consider the following optimization problem for linear regression $\mathbf{x}^i \in \mathbb{R}^d, y^i \in \mathbb{R}$. In the box below, write down a likelihood distribution for $\mathbb{P}[y^i \mid \mathbf{x}^i, \mathbf{w}]$ and prior $\mathbb{P}[\mathbf{w}]$ such that $\hat{\mathbf{w}}_{\text{rnc}}$ is the MAP estimate for your model. Give explicit forms for the density functions but you need not calculate normalization constants.

$$\hat{\mathbf{w}}_{\text{rnc}} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \sum_{i=1}^{n} (y^i - \langle \mathbf{w}, \mathbf{x}^i \rangle)^2 + \|\mathbf{w}\|_2^2$$
$$\text{s.t. } \|\mathbf{w}\|_2 \leq 1.$$

$\mathbb{P}[y^i \mid x^i, w] = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-(y^i - \langle w, x^i \rangle)^2 / 2\sigma^2\right)$ σ^2 is variance.

~~$\mathbb{P}[w] = \begin{cases} \frac{1}{\rho} & \|w\|_2 < 1 \rightarrow \text{uniform in unit sphere.} \\ 0 & \|w\|_2 > 1 \rightarrow \text{otherwise.} \end{cases}$~~

~~ρ is volume of the unit sphere in $\mathbb{R}^d$.~~

$\mathbb{P}[w] = \begin{cases} c\, e^{-\|w\|^2} & \|w\|_2 < 1 \\ 0 & \|w\|_2 > 1 \end{cases}$ c calculated by $c \int_0^{r} e^{-\frac{r^2}{}} = 1$ ~~ρ is variance~~

optionally we could take $e^{-\frac{\|w\|^2}{\rho^2}}$, ρ denoting shape/variation.

4. Recall that we derived the k-means algorithm by considering a Gaussian mixture model and forcibly setting the mixture proportions to $\pi_k^t = \frac{1}{K}$ as well as the covariance matrices of the Gaussians to identity $\Sigma^{k,t} = I$. Suppose we instead set $\Sigma^{k,t} = \Sigma$ where $\Sigma \in \mathbb{R}^{d \times d}$ is a known positive definite matrix. How will the k-means algorithm change due to this? Give the final algorithm below (no derivations required).

1. Initialise means $\{\mu^{k,0}\}_{k=1,\dots,K}$
2. for $i \in [n]$ update $z^{i,t}$ using $\mu^{k,t}, \Sigma$.
   $z^{i,t} = \arg\max_k \mathcal{N}(x^i \mid \mu^{k,t}, \Sigma)$
3. update $\mu^{k,t+1} = \arg\max_{\mu^{k,t}} \sum_{i: z^{i,t} = k} (x^i - \mu^{z_{i,t}})^T \Sigma^{-1} (x^i - \mu^{z_{i,t}})$
4. Repeat the process till convergence.
5.

translate to $\mu^{k,t+1} = \frac{1}{n_k^t} \sum_{i: z^{i,t} = k} x^i$ only.

Name: Raktim Mitra

Roll No.: 150562 Dept.: CSE

IIT Kanpur
CS771 Intro to ML
Mid-semester Examination
Date: September 21, 2017

Problem 4 (Long Answer: 3+3+5+5=16 marks). In this question we will derive an MLE estimate for a multinoulli distribution. Consider a K-faced die with faces $k = 1, 2, \ldots, K$. Let the vector $\boldsymbol{\pi}^*$ denote the vector encoding the probabilities of the various faces turning up i.e. face k turns up with probability $\boldsymbol{\pi}^*_k$. Clearly $\boldsymbol{\pi}^*_k \geq 0$ and $\sum_{k=1}^K \boldsymbol{\pi}^*_k = 1$. Now suppose I get n rolls of this die. Let $\mathbf{x} \in \mathbb{N}^K$ denote the vector that tells me how many times each face turned up i.e. the k-th face is found turning up $\mathbf{x}_k \geq 0$ times with $\sum_{k=1}^K \mathbf{x}_k = n$ (recall $\mathbb{N} = \{0, 1, 2, \ldots\}$ is the set of natural numbers). It turns out that we have $\mathbb{P}[\mathbf{x} \mid \boldsymbol{\pi}^*] = \frac{n!}{\prod_{k=1}^K (\mathbf{x}_k!)} \prod_{k=1}^K (\boldsymbol{\pi}^*_k)^{\mathbf{x}_k}$.

1. Write down the problem of finding the MLE estimate $\arg\max_{\boldsymbol{\pi}} \mathbb{P}[\mathbf{x} \mid \boldsymbol{\pi}]$ as an optimization problem. *Hint*: it will be a constrained optimization problem.

$$\hat{\pi}_{MLE} = \arg\max_{\pi} \sum_{k=1}^{K} -x_k \log(\pi_k) \quad \text{(after taking -log)}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} x_k = n, \quad \underbrace{\sum_{k=1}^{K} \pi_k = 1}_{\text{main constraint } \cancel{g(\pi)} = 1} \quad \pi_k \geq 0$$

2. Write down the Lagrangian for that optimization problem.

So, $\mathcal{L}(\pi, \alpha) = f(\pi) + \alpha g(\pi)$ $\quad f(\pi) = -\sum_{k=1}^{K} x_k \log(\pi_k)$ $\quad$ Let, $\sum_{k=1}^{K} \pi_k - 1 = g(\pi)$ $\Rightarrow$ Constraint $g(\pi) = 0$.

$g(\pi) = \sum_{k=1}^{K} \pi_k - 1$

$\alpha \in \mathbb{R}$.

3. Find the dual problem and eliminate the primal variable. Show major steps. Give the simplified dual problem which should be only in terms of constants and the dual variable.

Primal problem ~~$\hat{\pi}_{MLE} = \arg\max_{\pi}$~~ $\hat{\pi}_P = \arg\min_{\pi} \left\{ \arg\max_{\alpha \in \mathbb{R}} \left\{ \sum_{k=1}^{K} x_k \log(1/\pi_k) + \alpha \sum_{k=1}^{K} \pi_k - \alpha \right\} \right\}$

$\Rightarrow$ Dual problem: $\hat{\pi}_D = \arg\max_{\alpha \in \mathbb{R}} \left\{ \arg\min_{\pi} \left\{ \sum_{k=1}^{K} x_k \log(1/\pi_k) + \alpha \sum_{k=1}^{K} (\pi_k) - \alpha \right\} \right\}$

$L = \arg\min_{\pi} \sum_{k=1}^{K} x_k \log(1/\pi_k) + \alpha \sum_{k=1}^{K} \pi_k - \alpha$

$$\frac{\partial L}{\partial \pi_k} = x_k \pi_k \cdot -\frac{1}{\pi_k^2} + \alpha \Rightarrow -\frac{x_k}{\pi_k} + \alpha = 0 \Rightarrow x_k = \pi_k \cdot \alpha$$

$$\Rightarrow \pi_k = \frac{x_k}{\alpha}$$

So, removing π_k.

$$\hat{\pi}_D = \arg\max_{\alpha} \sum_{k=1}^{K} x_k \log(\alpha / x_k) + \sum_{k=1}^{K} \alpha \cdot \frac{x_k}{\alpha} - \alpha$$

$$= \arg\max_{\alpha} \sum_{k=1}^{K} x_k \log(\alpha / x_k) + (n - \alpha) \quad \sum_{k=1}^{K} x_k = n.$$

Name: Rakthim Mitra

Roll No.: 150562

Dept.: CSE

IIT Kanpur
CS771 Intro to ML
Mid-semester Examination
Date: September 21, 2017

4. Solve the dual problem and use it to obtain the MLE estimate. Only give expressions for both the dual solution as well as the MLE estimate.

$\alpha = \sum_{k=1}^{K} x_k = n \Rightarrow \alpha = \underline{n.}$ <u>Answer.</u>

$\hat{\pi} = \arg\min_{\pi} \sum_{k=1}^{K} x_k \log(1/\pi_k) + n \sum_{k=1}^{K} \pi_k - n.$

after solving, $-\frac{x_k}{\pi_k} + n = 0 \Rightarrow \pi_k = \frac{x_k}{n}.$

So, $\hat{\pi} = \left[\frac{x_1}{n}, \frac{x_2}{n}, \dots, \frac{x_K}{n}\right]$

—— <u>Answer</u>

BLANK SPACE: Any answers written here will be left ungraded.
No exceptions.
You may use this space for rough work.