# COMS 4771 Exam 1

Parita Pooj

TOTAL POINTS

## 74.5 / 100

QUESTION 1

### 1 Problem 1 (linear classifiers) 20 / 20

- **0 Correct pairs, (mostly) correct justification**
- **2** Correct pairs, some false statements in justification
- **3** Only one pair correct, one correct justification
- **7** No correct justifications
- **10** No justifications
- **16** Blank, or answer not even in correct form
- **0.25** Answer in wrong place without explanation
- **0.5** Difficult to read/find answer

QUESTION 2

### Problem 2 (features and kernels) 20 pts

### 2.1 Problem 2a (feature map) 10 / 10

- **0** Correct answer, (mostly) correct justification that mentions 1-to-1 correspondence of the linear separators.
- **0 Correct answer, weak/incomplete justification (e.g., 1-to-1 correspondence of data, which is trivial, or simply asserting the claim in different words)**
- **0** Correct answer, false statements or completely irrelevant statements (e.g., irrelevant stuff about kernels) in justification
- **0** Incorrect answer, or no justification
- **0** Blank, or not even a valid answer
- **0.25** Answer in wrong place without explanation
- **0.5** Difficult to read/find answer
- **0** Click here to replace this description.

### 2.2 Problem 2b (kernelized 1-NN) 4.5 / 10

- **0** Correct and unambiguous pseudocode
- **2** Pretends -K(x_i,x_new) is the distance function (i.e., forgets the +K(x_i,x_i))
- **4** Pretends K(x_i,x_new) is the distance function
- **8** Blank, or not even pseudocode
- **0.5 Ambiguity in pseudocode**
- **2.5** Assuming ψ linear
- **0.25** Answer in wrong place without explanation
- **0.5** Difficult to read/find answer
- **5 Other incorrect distance function, or no defined distance function beyond ||x-x'|| / ||ψ(x)-ψ(x')|| / "dist(x,y)"**
- **1** Missing or incorrect return value
- **0.5** Overly complex solution
- **3** Incorrect or incomplete algorithm
- **0.5** Maximizing instead of minimizing distance
- **2.5** Not using kernel to calculate ||x||^2, ||x_new||^2, or using kernel incorrectly despite correct mathematical derivation
- **0.5** Sign error

QUESTION 3

### Problem 3 (Naive Bayes) 30 pts

### 3.1 Problem 3a (error rates) 6 / 10

- **0** Correct answer, mostly correct justification via non-zero error rate of optimal (i.e., Bayes) classifier for $\hat P$. OK if reason is something like "because 0 < \hat\mu_{y,j} < 1 and \pi_y > 0 for all y,j"
- **1** Correct answer, ambiguous, irrelevant, or inane justification (e.g., statements about difference between training error and test error; simply restating the claim in different words)
- **2** Correct answer, false statements in justification
- **4 Incorrect answer, or no justification**
- **8** Blank, or not even a valid answer
- **0.25** Answer in wrong place without explanation
- **0.5** Difficult to read/find answer

### 3.2 Problem 3b (linear separability) 8.5 / 10

- **0** Correct answer, (mostly) correct justification by appealing to fact that zero training error rate on S is achieved by a linear classifier. Ok if just says S is

linearly separable, and thus S_{1,2} is, too. (No need to actually construct the linear separator for S_{1,2}.)

**- 1.5 Correct answer, ambiguous, irrelevant, or inane justification (e.g., simply restating the claim in different words)**

- 3 Correct answer, false statements in justification
- 4 Incorrect answer, or no justification
- 8 Blank, or not even valid answer
- 0.25 Answer in wrong place without explanation
- 0.5 Difficult to read/find answer

### 3.3 Problem 3c (linear classifiers) 6 / 10

- 0 Correct equation
- 0.1 Correct equation, but a lot of unnecessary extra stuff that anyways equals zero
- 0.5 Only gives an expression e such that "e = 0" or "e = 1" is a correct equation
- 1 Equation close to correct with slight mistakes (e.g. also contains some unnecessary stuff about the class priors)
- 2 Incorrect equation involving just \hat\mu_{y,j}'s

**- 4 Incorrect equation involving \hat\mu_{y,j}'s and other symbols including but not limited to \hat\pi_y's, d's, k's and x's**

- 8 Blank, or not even an equation in terms of parameters
- 0.25 Answer in wrong place without explanation
- 0.5 Difficult to read/find answer
- 1 Formula given with no justification
- 1 Solution is conceptually correct, but incorrect due to computational error
- 8 Not a relevant answer

QUESTION 4

## Problem 4 (convex optimization) 30 pts

### 4.1 Problem 4a (Hessian) 7 / 10

- 0 Correct Hessian. May define vector $u_i$ such that $\sum_{i=1}^{n/2} u_i u_i^T$ is the Hessian. (Some typos okay, e.g., n/2 is replaced by n.)
- 1 PSD matrix that is sum_{i=1}^{n/2} c_i x_i x_i^T, except c_i is an incorrect expression involving exp(y_i<w,x_i>)
- 2 Some scalar terms close to correct expression (just involving p_i := exp(y_i<w,x_i>/(1+exp(y_i<w,x_i>) and 1 - p_i), but incorrect matrix (e.g., a multiple of ww^T, or just diagonal matrix) or not a matrix

**- 3 A first derivative (gradient), or a far-from-correct second derivative**

- 8 Blank, or not even a valid answer
- 0.25 Answer in wrong place without explanation
- 0.5 Difficult to read/find answer

### 4.2 Problem 4b (problem convexity) 7 / 10

- 0 Correct answer, correct justification (should assert that both objective function and constraint functions are convex)
- 1 Correct answer, but fails to assert convexity of either objective function or constraint functions
- 2 Incorrect answer, but reason given is that constraint functions are not convex

**- 3 Incorrect answer (with incorrect or no justification), or no justification**

- 8 Blank, or not even valid answer
- 0.25 Answer in wrong place without explanation
- 0.5 Difficult to read/find answer

### 4.3 Problem 4c (new problem) 5.5 / 10

- 0 Correct optimization problem (okay if constraint is just "\|w\|_2 \leq 0"), correct answer, correct justification
- 8 Blank, or not even valid answers

**- 1.5 Missing new optimization problem**

- 1 Incorrect new optimization problem (e.g., incorrect objective function, incorrect constraint, additional constraints like those from part (a))

**- 1.5 Missing assessment of convexity of new optimization problem (or assessment of convexity without an optimization problem)**

- 1 Incorrect assessment of convexity of new optimization problem (e.g., problem is convex but answer is "no", problem is not convex but answer is "yes")

**- 1.5 Missing justification (or justification w/o an answer or optimization problem)**

- 1 Incorrect or ambiguous justification of answer (e.g.,

incorrect or missing reason for convexity/lack of convexity for either objective function or constraint function(s))

**- 0.25** Answer in wrong place without explanation

**- 0.5** Difficult to read/find answer

# Exam #1

COMS 4771 Machine Learning (Fall 2016)

Name: PARITA POOJ UNI: psp2133

**Instructions:** Write your name and UNI at the top of every page in the spaces provided, and agree to the pledge (below) by signing your name and writing today's date in the spaces provided. Failure to follow instructions will result in a failing grade.

Write your answers to the problems in the spaces provided (marked by "**Your answer**"). If you run out of room for an answer, continue on the back of the page. **Your answers should be precise, legible, and unambiguously separated from any other marks.** If it is difficult to find or read your answer, you will not receive any credit for it.

**Honor pledge:** I have not given or received unauthorized assistance on this examination. I will not retain or re-distribute any copies of this examination, either in part or in full.

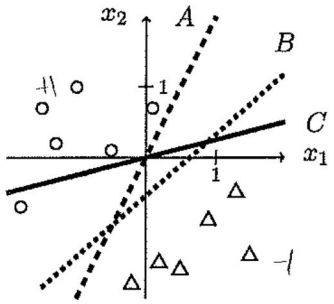
Sign here: Parita Date: 10/19/2016

(Do not write in the boxes below.)

| | |
|---|---|
| Problem 1 (20 points) | |
| Problem 2 (20 points) | |
| Problem 3 (30 points) | |
| Problem 4 (30 points) | |
| Total (100 points) | |

Name: PARITA POOJ UNI: psp2133

**Problem 1** (20 points). The figure below depicts the decision boundaries of three linear classifiers (labeled $A$, $B$, and $C$) and the locations of labeled training data (positive points are circles, negative points are triangles).




Consider the following algorithms for learning linear classifiers which could have produced the depicted classifiers given the depicted training data:

(1) ERM for homogeneous linear classifiers A

(2) Online Perceptron (making one pass over the training data) C

(3) An algorithm that exactly solves the SVM problem B

What is the most likely one-to-one correspondence between these algorithms and the depicted linear classifiers? *State your answer as a list of pairs* (e.g., $(1, A)$, $(2, B)$, $(3, C)$), and *briefly justify your answer.*

**Your answer (Problem 1):**

1) (1,A) since the classifier passes through the origin and does not misclassify any points in the sample.

(2,C) Since Online Perceptron can make a few mistakes after only the first pass.

(3, B) SVM solution maximizes the margin and finds the exact solution when one exists. For the given data points B fits these criteria perfectly.

Name: PARITA POOJ UNI: psp2133

**Problem 2** (10+10=20 points).

(a) Assume the data set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ from $\mathbb{R}^d \times \{-1, +1\}$ is linearly separable, and let $\hat{f}$ be the SVM classifier on this data set. Also, assume the matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ has rank $d$, and define a feature transformation $\boldsymbol{\phi}: \mathbb{R}^d \to \mathbb{R}^d$ by $\boldsymbol{\phi}(\boldsymbol{x}) := \boldsymbol{A}\boldsymbol{x}$. Does the SVM classifier $\hat{f}^\sharp$ obtained from the transformed data set $\{(\boldsymbol{\phi}(\boldsymbol{x}_i), y_i)\}_{i=1}^n$ satisfy

$$\hat{f}^\sharp(\boldsymbol{\phi}(\boldsymbol{x})) \;=\; \hat{f}(\boldsymbol{x}) \quad \text{for all } \boldsymbol{x} \in \mathbb{R}^d\,?$$

*Answer with either "yes" or "no", and briefly justify your answer.*

(b) Let $K: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a kernel that corresponds to a feature map $\boldsymbol{\psi}: \mathbb{R}^d \to \mathbb{R}^D$. Your goal is to construct a 1-nearest neighbor classifier, based on training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ from $\mathbb{R}^d \times \{0, 1\}$, where the distance between any two points $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$ is $\|\boldsymbol{\psi}(\boldsymbol{x}) - \boldsymbol{\psi}(\boldsymbol{x}')\|_2$. Assume you have a subroutine for computing $K(\boldsymbol{x}, \boldsymbol{x}')$ for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$, but you do not have a subroutine for computing $\boldsymbol{\psi}(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathbb{R}^d$. *Write precise and unambiguous pseudocode* for computing the 1-nearest neighbor prediction for a new input point $\boldsymbol{x}_{\text{new}} \in \mathbb{R}^d$.

**Your answer (Problem 2a):** (Yes)

$\phi(x) = Ax$ for an invertible A matrix corresponds to linear transformation of $x$. Since, SVM has a unique solution when a linear classifier exists $\hat{f}^{\#}(\phi(x)) = \hat{f}^{\#}(Ax) = \hat{f}(x)$ for all $x \in R^d$

Name: PARITA POOJ UNI: psp2133

**Your answer (Problem 2b):**

$K(x, x')$

$$\|\psi(x) - \psi(x')\|_2^2 = \|\psi(x)\|_2^2 - 2\|\langle\psi(x), \psi(x')\|_2^2 + \|\psi(x)\|_2^2$$

$$\therefore \|\psi(x) - \psi(x')\|_2^2 = \|\psi(x)\|_2^2 + \|\psi(x)\|_2^2 - 2\|\sum \psi(x) K(x, x')\|_2 \quad \text{---- (1)}$$

† Using the above formula, we can get

1) For $x_{new} \in R^q$, ~~compute $\psi$~~

compute $\|\psi(x_{new}) - \psi(x)\|_2^2$

we need $\|\psi(x)\|_2^2 \rightarrow$ previously stored

$\|\psi(x_{new})\|_2^2$ compute

& $\|\psi(x) K(x, x_{new})\|_2^2$

~~This~~

2) Compare this dist based on (1) for all $x$

3) Choose $x$ with lowest dist.

Name: PARITA POOJ UNI: psp2133

**Problem 3** (10+10+10=30 points). Recall that in the Bernoulli Naïve Bayes generative model for the $K$-class classification problem, the class conditional distributions are of the form $P_{\mu_y}(\boldsymbol{x}) = \prod_{j=1}^{d} \mu_{y,j}^{x_j}(1-\mu_{y,j})^{1-x_j}$ for $\boldsymbol{x} = (x_1, x_2, \ldots, x_d) \in \{0,1\}^d$. Here, $\boldsymbol{\mu}_y = (\mu_{y,1}, \mu_{y,2}, \ldots, \mu_{y,d}) \in [0,1]^d$ is the parameter vector for class $y \in \{1, 2, \ldots, K\}$ from the parameter space $[0,1]^d$. The class priors are denoted by $\pi_1, \pi_2, \ldots, \pi_K$.

(a) Training data $S := \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ from $\{0,1\}^d \times \{1, 2, \ldots, K\}$ is used to estimate parameters of the Bernoulli Naïve Bayes generative model, using the approach from the homework assignment. The estimated parameters are denoted by $(\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_K, \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \ldots, \hat{\boldsymbol{\mu}}_K)$, and the corresponding distribution from the generative model is denoted by $\widehat{P}$.

Suppose the estimated parameters satisfy $\hat{\pi}_y = 1/K$ and $0 < \hat{\mu}_{y,j} < 1$ for all $j \in \{1, 2, \ldots, d\}$ and $y \in \{1, 2, \ldots, K\}$; and furthermore, suppose the Bayes classifier $\hat{f}$ for $\widehat{P}$ has training error rate $\mathrm{err}(\hat{f}, S) = 0$. Is it possible to have $\mathrm{err}(\hat{f}, \widehat{P}) = 0$ (i.e., for the error rate of $\hat{f}$ with respect to $\widehat{P}$ to be zero)? *Answer with either "yes" or "no", and briefly justify your answer.*

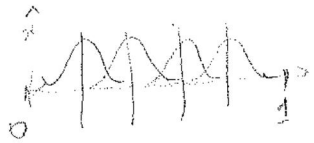
(b) Continuing from (a), let $S_{1,2}$ be the data set comprised only of examples $(\boldsymbol{x}_i, y_i)$ from $S$ such that $y_i \in \{1, 2\}$. Is the data set $S_{1,2}$ linearly separable? *Answer with either "yes" or "no", and briefly justify your answer.*

(c) Continuing from (a) and (b): now, $S_{1,2}$ is used to estimate parameters $\tilde{\boldsymbol{\theta}} := (\tilde{\pi}_1, \tilde{\pi}_2, \tilde{\boldsymbol{\mu}}_1, \tilde{\boldsymbol{\mu}}_2)$ of the same generative model, except with just two classes, using the same method from the homework assignment. Let $\widetilde{P}$ denote the corresponding distribution from the generative model. Determine an equation (possibly non-linear) that is satisfied by the parameters $\tilde{\boldsymbol{\mu}}_1$ and $\tilde{\boldsymbol{\mu}}_2$ if and only if the Bayes classifier for $\widetilde{P}$ is a homogeneous linear classifier. (Recall that a homogeneous linear classifier is a linear classifier with a decision boundary that passes through the origin.) *Write the equation explicitly* (e.g., "$\sum_{j=1}^{d} \tilde{\mu}_{1,j}\tilde{\mu}_{2,j} = 0$"), and *briefly justify your answer.*

**Your answer (Problem 3a):** (Yes)

Yes, it is possible to have $\mathrm{err}(\hat{f}, \hat{P}) = 0$ if the data is from an i.i.d. sample. This is possible when the generative model perfectly specifies the probability distribution for $x$.

Example case can be considered where the predicted gaussians perfectly represent the distribution for the entire $\forall x \in R^d$.



Name: PARITA POOJ UNI: psp2133

**Your answer (Problem 3b):** (Yes)

for only two labels 1,2 :

$$\hat{f}: x \longmapsto \underset{y \in \{1,2\}}{\arg\max}\ \pi_y p_{i|y}$$

$$= \underset{y \in \{1,2\}}{\arg\max} \left( \hat{\pi}_y \prod_{j=1}^{d} \mu_{y,j}^{x_j} (1-\mu_{y,j})^{1-x_j} \right)$$

$$\ell(x; \mu, \pi) = \underset{y \in \{1,2\}}{\arg\max} \Big[ \log \hat{\pi}_y + \sum_{j=1}^{d} \big[ x_j \log \mu_{y,j} + (1-x_j) \log(1-\mu_{y,j}) \big]$$

From above we can simplify and get a linear classifier $\langle \alpha, x \rangle + \alpha_0 = 0$ as we did in class and homework. Thus, we can have

$$f: x \mapsto \begin{cases} 1 & \langle \alpha, x \rangle + \alpha_0 > 0 \\ 2 & \langle \alpha, x \rangle + \alpha_0 \leq 0 \end{cases}$$

Hence, linearly separable.

**Your answer (Problem 3c):**

Bayes classifier when it is a homogeneous linear classifier has

$$f(x) = \begin{cases} 1 & P(Y=1 \mid X=x) > P(Y=2 \mid X=x) \\ 2 & P(Y=2 \mid X=x) > P(Y=1 \mid X=x) \end{cases}$$

i.e. $f(x) = P(Y=1 \mid X=x) - P(Y=2 \mid X=x) - 1 > 0$

~~$f(x) =$~~ $P(Y=1 \mid X=x) = P(Y=2 \mid X=x)$

$P(Y=1 \mid X=x) = P(Y=2 \mid X=x) + 1$

$$\frac{\tilde{\pi}_1}{\tilde{\pi}_2} \prod_{j=1}^{d} \frac{\mu_{j,1}^{x_j} (1-\mu_{j,1})^{1-x_j}}{1 + \mu_{j,2}^{x_j} (1-\mu_{j,2})^{1-x_j}} = 0$$

# Exam #1

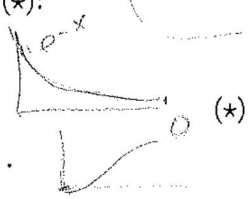
COMS 4771 Machine Learning (Fall 2016)

Name: PARITA ROUT UNI: psp2133

**Problem 4** (10+10+10=30 points). Suppose we have training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ from $\mathbb{R}^d \times \{-1, +1\}$ (where $n$ is an even positive integer). Consider the following optimization problem, which we refer to as $(\star)$:

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} \sum_{i=1}^{n/2} \ln\bigl(1 + \exp(-y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle)\bigr) \qquad (\star)$$

$$\text{s.t.} \quad \exp(-y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle) - 1/2 \leq 0 \quad \text{for all } i \in \{n/2+1, n/2+2, \ldots, n\}.$$




(a) *Write the Hessian matrix* (i.e., $d \times d$ matrix of second-derivatives) of the objective function in $(\star)$ at an arbitrary point $\boldsymbol{w} \in \mathbb{R}^d$. The matrix should be expressed in terms of $\boldsymbol{w}$ and the training data.

*Hint*: It should be possible to write the matrix as the sum of matrices of the form $\boldsymbol{u}\boldsymbol{u}^\top$, where $\boldsymbol{u}$ is a vector in $\mathbb{R}^d$ (regarded as a $d \times 1$ column vector). Matrices of this form are always symmetric and positive semidefinite.

(b) Is the optimization problem $(\star)$ convex? *Answer with either "yes" or "no", and briefly justify your answer.*

(c) *Write a new optimization problem*, in the standard form described in lecture, whose solution minimizes the objective function in $(\star)$ over all vectors in $\mathbb{R}^d$ with Euclidean length at most 10. Is this new problem convex? *Answer with either "yes" or "no", and briefly justify your answer.*

**Your answer (Problem 4a):**

The above can be transformed to minimizing loss function

$$L = \sum_{i=1}^{n/2} \ln(1 + \exp(-y_i \langle w, x_i \rangle)) + \sum_{i=n/2}^{n} \ln(\exp(-y_i \langle w, x_i \rangle - 1/2)$$

Let $\exp(-y_i \langle w, x_i \rangle) = z$

$$= \sum_{i=1}^{n/2} \sum_{i=1}^{n} \ln(1+z) - \sum_{i=n/2}^{n} \ln(1+z)$$

$$+ \sum_{i=1}^{n} \ln(z - 1/2) - \sum_{i=1}^{n/2} \ln(z - 1/2)$$

min this

$$= \sum_{i=1}^{n} \ln\bigl((1+z)(z-1/2)\bigr) - \sum_{i=1}^{n/2} \ln(z-1/2) - \sum_{i=n/2}^{n} \ln(1+z)$$

min

$$= \sum_{i=1}^{n} \ln\, z - 1/2 + z^2 - z/2$$

$$= \sum_{i=1}^{n} \ln(z^2 - z/2 - 1/2)$$

$$z^2 = e^{+ y_i^2 \langle w, x_i \rangle \langle w, x_i \rangle}$$

$$= e^{w^T x_i^T x_i w}$$

$$\frac{\partial L}{\partial x} = \sum_{i=1}^{n} w^T w \, x_i \qquad \frac{\partial L}{\partial x^2} = \sum_{i=1}^{n} w^T w$$

Name: PARITA POOJ UNI: psp2133

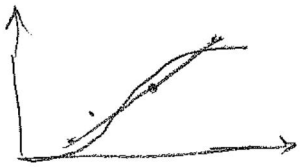
**Your answer (Problem 4b):** ~~(Yes)~~ (No)

The given function can be transformed and written in the form similar to a logistic regression function.

Let $z = y_i \langle w, x_i \rangle$

$$\min_{w \in R^d} \sum_{i=1}^{n/2} \ln($$

The function can be drawn as



which is not a convex function

**Your answer (Problem 4c):** Yes

$\sum$ can be written from (a) with $\|x\|_2^2 \leq 10$ for 1

Name: PARITA POOJ UNI: pjp2133

**This page is intentionally (mostly) blank. Use for any scratchwork/calculations. Will not be graded.**

Name: PARTH POOJ UNI: psp2133

**This page is intentionally (mostly) blank. Use for any scratchwork/calculations. Will not be graded.**