CS M146 Final Exam

MARK GUEVARA

TOTAL POINTS

85.5 / 118

QUESTION 1

1 True or False 15 / 15

✓ **- 0 pts Correct**

- 3 pts a) wrong

- 3 pts b) wrong

- 3 pts c) wrong

- 3 pts d) wrong

- 3 pts e) wrong

QUESTION 2

Naive Bayes 14 pts

2.1 (a)-(e) 8 / 14

- 0 pts Correct

- 3 pts a) wrong

- 1.5 pts a) partial

- 3 pts b) wrong

✓ **- 3 pts c) wrong**

✓ **- 3 pts d) wrong**

- 1 pts d) 1 wrong

- 2 pts d) 2 wrong

- 2 pts e) wrong

- 1 pts e) partial

QUESTION 3

Expectation Maximization 18 pts

3.1 (a)-(b) 6.5 / 9

- 0 pts Correct

- 2 pts a) partially correct

- 4 pts a) incorrect

✓ **- 2.5 pts b) partially incorrect**

- 5 pts b) incorrect

3.2 (c)-(d) 8 / 9

- 4 pts (c) incorrect (do not decompose P(x) into \sum_y P(x,y))

- 2 pts (c) minor mistake

- 5 pts (d) incorrect

- 3 pts (d) mention it is an iterative process. but the process is incorrect or is unclear

- 1 pts (d) mention alternative update P(Y|x; \theta) and parameter s but do not give sufficient details

- 5 pts (d) Incorrect

✓ **- 1 pts (d) missing some minor details**

- 0 pts Correct

QUESTION 4

Kernels and SVM 26 pts

4.1 (a) i Circle correct option 2 / 2

✓ **- 0 pts Correct**

- 2 pts Wrong

4.2 (a) ii Prove 1 2 / 5

- 0 pts Correct

✓ **- 3 pts mentioned k(x,x') = φ (x) φ^T (x') or linear combination, but did not mention φ' (x) = √c φ (x) or positive semi-definite**

- 5 pts incorrect

- 1 pts tiny mistake

4.3 (a) ii Prove 2 2 / 5

✓ **- 3 pts Only mentioned k(x,x') = φ(x)·φ(x') or linear combination**

- 2 pts Gave the feature mappings: φ(x) =< f1(x),f2(x),...,fn(x)>

- 0 pts Correct (Gave φ3(x) =< f1(x),f2(x),...,fn(x),g1(x),g2(x),...,gn(x)> or mentioned the concatenation (NOT numerical addition) of feature mappings or positive semi-definite)

- 5 pts Wrong

4.4 (b) i Unconstrained 0 / 4

- **2 pts** Wrong but gave the correct hinge loss
- **0 pts** Correct
- ✓ **- 4 pts** **Wrong**

4.5 (b) ii Remove constraints 3 / 6

- **0 pts** correct
- ✓ **- 3 pts** **partly correct**
- **6 pts** wrong
- **0 pts** Click here to replace this description.

4.6 (b) iii Support vectors 2 / 4

- **0 pts** Correct
- ✓ **- 2 pts** **mistake in Q1 (no partial credits)**
- **2 pts** mistake in Q2 (no partial credits)

QUESTION 5

PAC learning and VC dimension 15 pts

5.1 vc dim 6 / 6

- ✓ **- 0 pts** **Correct**
- **3 pts** wrong vc with proof
- **6 pts** wrong VC, no proof
- **1.5 pts** wrong e
- **1.5 pts** wrong delta
- **1.5 pts** wrong m
- **2.5 pts** partially correct math
- **5 pts** wrong ans
- **3 pts** wrong proof
- **1 pts** Incomplete proof: You need to show that VC can not be more than 2
- **4 pts** no proof
- **0.5 pts** No details proof
- **5 pts** You are supposed to find a numeric value and prove it.
- **6 pts** no ans for vc dim

5.2 PAC define 5 / 5

- ✓ **- 0 pts** **Correct**
- **1.5 pts** wrong/no definition of e
- **1.5 pts** wrong/no definition of delta
- **1.5 pts** wrong/no definition of m/inequality
- **0.5 pts** if error <=e, probability = 1-dela
- **0.5 pts** Not equal to e, correct ans: <=e
- **1 pts** Uncertainty that the error will be <= e is (1-delta)
- **1 pts** e means the max limit of error of any hypothesis in H
- **0 pts** Not less than e, correct ans: <=e
- **1.5 pts** e, and delta are swapped and error is not on the training data.
- **0.5 pts** partially correct explanation of #training example or sample complexity m or inequality
- **1 pts** Not equal to 1/e, correct ans: <=e; not equal to 1/delta, correct ans>=(1-delta)
- **0 pts** m is #training example/data

5.3 Math 4 / 4

- ✓ **- 0 pts** **Correct**
- **4 pts** Totally incorrect
- **2 pts** partially correct
- **0.5 pts** Wrong final ans
- **1 pts** without reasoning
- **0 pts** I don't understand what is your ans?

QUESTION 6

Short Answer Question 30 pts

6.1 (a)-(g) 22 / 30

- **0 pts** Correct
- **2 pts** a) wrong
- ✓ **- 2 pts** **b) partially wrong**
- **4 pts** b) wrong
- ✓ **- 3 pts** **c) partially wrong (e.g., mistake in derivation, say \theta = sum y_i).**
- **6 pts** c) wrong
- **3 pts** d) partially wrong
- **6 pts** d) wrong
- **2 pts** e) wrong
- ✓ **- 3 pts** **f) partially wrong**
- **5 pts** f) wrong
- **2 pts** g) partially wrong (e.g., did not mention prior or detailed conditions)
- **5 pts** g) wrong

gradescope

CM146: Introduction to Machine Learning — Fall 2018

Final Exam

- This is a closed book exam. Everything you need in order to solve the problems is supplied in the body of this exam.
- This exam booklet contains **six** problems.
- You have 150 minutes to earn a total of 120 points.
- **Besides giving the correct answer, being concise and clear is very important. To get the full credit, you must show your work and explain your answers.**

Good Luck!

Name and ID: Mark Guevara 704962920

Name and ID		/2
True/False		/15
Naive Bayes		/14
Expectation Maximization		/18
Kernels and SVM		/26
Learning Theory		/15
Short Answer Questions		/30
Total		/120

1 True or False [15 pts]

Choose either True or False for each of the following statements. (If your answer is incorrect, partial point may be given if your explanation is reasonable.)

(a) **(3 pts)** After mapping feature into a high dimension space with a proper kernel, a Perceptron may be able to achieve better classification performance on instances it wasn't able to classify before.

True

(b) **(3 pts)** Given two classifiers A and B, if A has a lower VC-dimension than B, then conceptually A is more likely to overfit the data.

False; A lower VC-dimension means a simpler model, and simpler models are less prone to overfitting.

(c) **(3 pts)** If two variables, X, Y are conditional independent given Z, then the variables X, Y are independent as well.

False; $P(X|Z)$ and $P(Y|Z)$ may be independent, but $P(X)$ and $P(Y)$ may not;

(d) **(3 pts)** The SGD algorithm for soft-SVM optimization problem does not converge if the training samples are not linearly separable.

False, soft-SVM can converge

(e) **(3 pts)** In the AdaBoost algorithm, at each iteration, we increase the weight for misclassified examples.

True, the weights of correct classifications are proportionally decreased

2 Naive Bayes [14 pts]

Imagine you are given the following set of training examples. Each feature can take one of three nominal values: a, b, or c.

F1	F2	F3	C
a	c	a	+
c	a	c	+
a	a	c	-
b	c	a	-
c	c	b	-

(a) **(3 pts)** What modeling assumptions does the Naive Bayes model make?

Naive Bayes assumes that every parameter is conditionally independent. In this case, $P(F1 \mid C)$, $P(F2 \mid C)$, and $P(F3 \mid C)$ are assumed to be independent.

(b) **(3 pts)** How many parameters do we need to learn from the data in the model you defined in (a).

$P(F1 = a \mid C = +)$
$P(F2 = a \mid C = +)$
$P(F3 = a \mid C = +)$
$\vdots$
$P(C)$

3 features · 3 classifications · 2 C + P(C)
F1, F2, F3 a, b, c +, -
= 19 parameters

(c) **(3 pts)** How many parameters do we have to estimate if we do not make the assumption as in Naive Bayes?

Many more than (6);
Every variable must be conditioned on every other variable

$P(F1 = a \mid F2, F3, C)$ ...

1 total
F1: a, b, c
F2
F3

3 (= a, = b, = c) · 3 (a, b, c) · 2 (+, -)

$9 \cdot 3 \cdot 3 \cdot 2 + 1$
$= 81 \cdot 2 + 1$
$= 163$

$P(C)$

$P(C)$

(d) **(3 pts)** We use the maximum likelihood principle to estimate the model parameters in (a). Show the numerical solutions of any three of the model parameters.

$P(F1 = a \mid C = +) = 1/2$

$P(F1 = b \mid C = +) = 0$

$P(F1 = c \mid C = +) = 1/2$

(e) **(2 pts)** In two sentences, describe the difference between logistic regression and the Naive Bayes?

Logistic regression uses a sigmoid function to find an optimal distribution for the data. Naive Bayes uses the existing probabilities of the data to assume an optimal distribution.

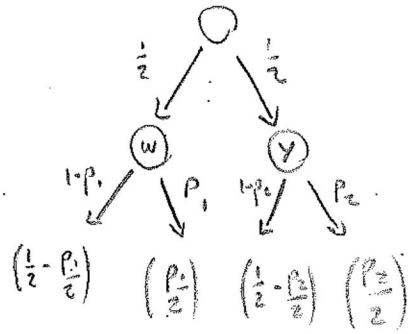
Logistic regression is also discriminative while Naive Bayes is generative.

3 Expectation Maximization [18 pts]

Suppose that somebody gave you a bag with two biased coins, having the probability of getting heads of p_1 and p_2 respectively. You are supposed to figure out p_1 and p_2 by tossing the coins repeatedly. You will repeat the following experiment n times: pick a coin uniformly at random from the bag (i.e. each coin has probability $\frac{1}{2}$ of being picked) and toss it, recording the outcome. The coin is then returned to the bag. Assume that each experiment is independent.

(a) **(4 pts)** Suppose the two coins have different colors: the white coin has probability p_1 to show head, while the yellow coin has probability p_2 to show head. Based on color, you know which coin you tossed during the experiments. After n tosses, the numbers of heads and tails of the white coin are H_1 and T_1, respectively. And, the number of heads and tails of the yellow coin are H_2 and T_2. Write down the likelihood function.

$$L = p_1^{H_1}(1-p_1)^{T_1} p_2^{H_2}(1-p_2)^{T_2}$$



(b) **(5 pts)** Based on the above likelihood function. Derive the maximum likelihood estimators for the two parameters p_1 and p_2 (please provide the detailed derivations).

$$\frac{\partial L}{\partial p_1} = -H_1 p_1^{(H_1-1)} \cdot T_1 (1-p_1)^{T_1-1} p_2^{H_2} (1-p_2)^{T_2}$$

$$\frac{\partial L}{\partial p_2} = -p_1^{H_1}(1-p_1)^{T_1} H_2 p_2^{H_2-1} \cdot T_2 (1-p_2)^{T_2-1}$$

both $= 0$, set equal,

$$\frac{\partial L}{\partial p_1} = \frac{\partial L}{\partial p_2}$$

$$\frac{H_1}{p_1} \cdot \frac{T_1}{(1-p_1)} \cdot \frac{p_2}{H_2} \cdot \frac{(1-p_2)}{T_2} = 1 \Leftarrow \frac{\frac{\partial L}{\partial p_1}}{\frac{\partial L}{\partial p_2}} = 1$$

$$\Rightarrow \frac{H_1 T_1}{H_2 T_2} p_2 (1-p_2) = p_1 (1-p_1) \Rightarrow \text{Solve for } p_1, p_2 ?$$

(c) **(4 pts)** Now suppose both coins look identical, hence the identity of the coin is missing in your data. After n tosses, if the numbers of heads and tails we got are H and T, respectively. Write down the likelihood function. Describing the challenge of maximizing this likelihood function.

Since we do not know which coin corresponds to which flips, we must make a lot of assumptions to find the most likely solution.

$$L = (p_1 + p_2)^H \cdot \left((1-p_1) + (1-p_2)\right)^T$$

(d) **(5 pts)** Describe how to optimize the likelihood function in the previous question by the EM algorithm (please provide sufficient details).

1. First select some initial assumption about the distribution of the two coins, θ (i.e. which flips correspond to which coin)
2. Calculate the expected values of p_1 and p_2 using that θ.
3. With the new p_1 and p_2, find the expected value of θ.
4. Repeat steps 2 and 3 until convergence.

4 Kernels and SVM [26 pts]

(a) In this question we will define kernels, study some of their properties and develop one specific kernel.

i. **(2 pts)** Circle the correct option below:
A function $K(x, z)$ is a valid kernel if it corresponds to the _____ {"inner(dot) product", "sum"} in some feature space, of the feature representations that correspond to x and z.

$K(x, z) = \phi(x)^T \phi(z)$

ii. **(10 pts)** In the next few questions we guide you to prove the following properties of kernels:
Linear Combination Property: if $\forall\, i, k_i(x, x')$ are valid kernels, and $c_i > 0$ are constants, then $k(x, x') = \sum_i c_i k_i(x, x')$ is also a valid kernel.

- **(5 pts)** Given a valid kernel $k_1(x, x')$ and a constant $c > 0$, use the definition above to show that $k(x, x') = ck_1(x, x')$ is also a valid kernel.

$k(x, x') = c k_1(x, x') = \sum_{i=1}^{1} c_i k_i(x, x')$, where $c_1 = c$

$\Rightarrow k(x, x')$ is a valid kernel

- **(5 pts)** Given valid kernels $k_1(x, x')$ and $k_2(x, x')$, use the definition above to show that $k(x, x') = k_1(x, x') + k_2(x, x')$ is also a valid kernel.

$k(x, x') = k_1(x, x') + k_2(x, x')$

$= \sum_{i=1}^{2} c_i k_i(x, x')$

where $c_1, c_2 = 1$

$\Rightarrow k(x, x') = \sum_{i=1}^{2} k_i(x, x') \Rightarrow k$ is a valid kernel

(b) Let $\{(x_i, y_i)\}_{i=1}^{l}$ be a set of l training pairs of feature vectors and labels. We consider binary classification, and assume $y_i \in \{-1, +1\} \forall i$. The following is the primal formulation of L2 SVM, a variant of the standard SVM obtained by squaring the hinge loss:

$$\begin{aligned} \min_{w,\xi_i,b} \quad & \frac{1}{2} w^T w + \frac{C}{2} \sum_i \xi_i^2 \\ \text{s.t} \quad & y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \forall i \in \{1, ..., l\} \\ & \xi_i \geq 0, \quad \forall i \in \{1, ..., l\} \end{aligned} \tag{1}$$

i. **(4 pts)** Derive an unconstrained optimization problem that is equivalent to Eq. (1).

$$\min_{w,\xi_i,b} \quad \frac{1}{2} w^T w + \frac{C}{2} \sum \xi_i^2$$

$$\text{s.t.} \quad y_i(w^T x + b) \geq 1 - |\xi_i|, \quad \forall i \in \{1, ..., l\}$$

ii. **(6 pts)** Show that removing the last set of constraints $\{\xi \geq 0, \forall i\}$ does not change the optimal solution to the primal problem.

ξ_i is only used in two places in the formulation: the summation $\sum_i \xi_i^2$ and the inequality, in $1 - \xi_i$. ξ_i^2 is always positive, so the constraint will not matter and ξ_i can be less than 0. Adding an abs. val to the inequality means $\xi_i < 0$ is functionally the same as its positive counterpart, so the actual solution will be unchanged.

iii. **(4 pts)** Given the following dataset in 1-d space, which consists of 4 positive data points $\{0, 1, 2, 3\}$ and 3 negative data points $\{-3, -2, -1\}$.

- if $C = 0$, please list all the support vectors.

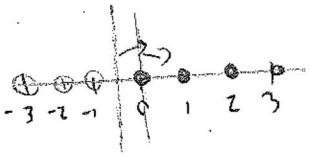
No margin

$\min \frac{1}{2} w^T w \quad \text{s.t.} \quad y_i(w^T x + b) \geq 1 - \xi_i$

Let $\xi_i = 1 \Rightarrow y_i(w \cdot x + 0) \geq 0 \Rightarrow w = 0$

$b = 0$

0



- if $C \to \infty$, please list all the support vectors.

Hard SVM

-1, 0

huge penalty, so want $\xi = 0$

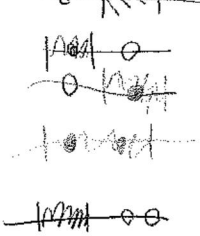
$y(w \cdot x + b) \geq 1$

5 PAC learning and VC dimension [15 pts]

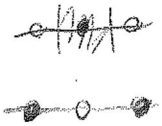
(a) **(6 pts)** Consider a learning problem in which $x \in R$ is a real number, and the hypothesis space is a set of intervals $H = \{(a < x < b) | a, b \in R\}$. Note that the hypothesis labels points inside the interval as positive, and negative otherwise. What is the VC dimension of H?

VC = 1

VC = 2



VC = 3



Not shatterable

● = + example
○ = - example

Since the x values must be in line with each other ($x \in R$) and the hypothesis only labels points inside the region, VC = 3 cannot be shattered.

$\Rightarrow$ **VC(H) = 2**

(b) **(5 pts)** The sample complexity of a PAC-learnable Hypothesis class H is given by

$$m \geq \frac{\log(|H|/\delta)}{\epsilon} \tag{2}$$

In three sentences, explain the meaning of ϵ and δ and the meaning of the inequality.

ϵ is the error and δ is the confidence interval. If you want a smaller error, you will need a greater minimum number of samples to lower that error. If you want to be more confident in your answer, δ goes down, since $(1-\delta)$ is the percent confidence, so more samples are needed.

(c) **(4 pts)** Now suppose we have a training set with 25 examples and our model has an error of 0.32 on the test-set. Based on Eq. (2), how many training examples we may need to reduce the error rate to 0.15 ? (Only need to list the formulation.)

$$25 \geq \frac{1}{0.32} \log\left(\frac{|H|}{\delta}\right)$$

$$\Rightarrow \log\left(\frac{|H|}{\delta}\right) \leq (0.32)(25)$$

$$m \geq \frac{1}{0.15} \log\left(\frac{|H|}{\delta}\right)$$

$$m \geq \frac{(0.32)(25)}{0.15}$$

≈ 50 (a little more)

6 Short Answer Questions [30 pts]

Most of the following questions can be answered in one or two sentences. Please make your answer concise and to the point.

(a) **(2 pts)** Multiple choice: for a neural network, which one of the following design choices that affects the trade-off between underfitting and overfitting the most:

i. The learning rate

ii. The number of hidden nodes

iii. The initialization of model weights

(b) **(4 pts)** Describe the difference between *maximum likelihood* (MLE) and maximum a posteriori (MAP) principles, and under what condition, MAP is reduced to MLE?

MLE tries to learn $\arg\max P(X|h)$ and makes no assumptions about the hypothesis. MAP tries to learn $\arg\max P(X|h)P(h)$ and uses existing information to assume properties of the hypothesis. MAP is reduced to MLE when no prior information can be used and only the data set is available.

(c) **(6 pts)** If a data point y follows the Poisson distribution with rate parameter θ, then the probability of a single observation y is given by

$$p(y;\theta) = \frac{\theta^y \exp^{-\theta}}{y!}$$

You are given data points $y_1, y_2, ..., y_n$ independently drawn from a Poisson distribution with parameter θ. What is the MLE of θ. (Hint: write down the log-likelihood as a function of θ.)

$Y = \{y_1, \dots, y_n\}$ $\max_\theta P(Y|\theta) = \max_\theta \prod_{i=1}^{n} P(y_i;\theta) = \max_\theta \prod_{i=1}^{n} \frac{\theta^{y_i} e^{-\theta}}{y_i!}$

$$= \max_\theta \sum_{i=1}^{n} \ln\left(\frac{\theta^{y_i} e^{-\theta}}{y_i!}\right)$$

$$\ln P(Y;\theta) = \max_\theta \sum_{i=1}^{n} \ln(\theta^{y_i}) - \ln(y_i!) - \theta$$

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix}$$

(d) **(6 pts)** Given vectors x and z in R^3, define the kernel $K_\beta(x;z) = (\beta + x \cdot z)^2$ for any value $\beta > 0$. Find the corresponding feature map $\phi_\beta(\cdot)$.

$$(\beta + x \cdot z)^2 = (\beta + x_1 z_1 + x_2 z_2 + x_3 z_3)^2$$

$$= \beta^2 + 2\beta x_1 z_1 + 2\beta x_2 z_2 + 2\beta x_3 z_3 + 2x_1 x_2 z_1 z_2 + 2x_1 x_3 z_1 z_3$$
$$+ 2x_2 x_3 z_2 z_3 + x_1^2 z_1^2 + x_2^2 z_2^2 + x_3^2 z_3^2 = \phi(x)^T \phi(z)$$

$$\Rightarrow \boxed{\phi_\beta^T(x) = \left[\beta, \sqrt{2\beta}x_1, \sqrt{2\beta}x_2, \sqrt{2\beta}x_3, \sqrt{2}x_1x_2, \sqrt{2}x_1x_3, \sqrt{2}x_2x_3, x_1^2, x_2^2, x_3^2\right]}$$

(e) **(2 pts)** Your billionaire friend needs your help. She needs to classify job applications into good/bad categories, and also to detect job applicants who lie in their applications using density estimation to detect outliers. To meet these needs, do you recommend using a discriminative or generative classifier? Why?

Detecting outliers with density estimation means a probabilistic model is more useful, so a generative classifier may be more useful in most cases. If she already has existing data on good, bad, and lying applicants, a discriminative model may prove to be more useful, but since that is not stated a generative classifier will aim to create those categories.

(f) **(5 pts)** We consider Boolean functions in the class $L_{10,30,100}$. This is the class of 10 out of 30 out of 100, defined over $\{x_1, x_2, \ldots x_{100}\}$. Recall that a function in the class $L_{10,30,100}$ is defined by a set of 30 relevant variables. An example $x \in \{0,1\}^{100}$ is positive if and only if at least 10 out these 30 variables are on. In the following discussion, for the sake of simplicity, whenever we consider a member in $L_{10,30,100}$, we will consider the function f in which the first 30 coordinates are the relevant coordinates. Show a linear threshold function h that behaves just like $f \in L_{10,30,100}$ on $\{0,1\}^{100}$.

let $h = \begin{cases} 1 & h^T x \geq 0 \\ 0 & \text{otherwise} \end{cases}$, where $h^T = [h_1, h_2, \ldots h_{100}, b]$ ← bias term

Then $h^T = [h_1, \ldots h_{30}, h_{31}, \ldots, h_{100}, b]$

$h^T = [\frac{1}{30}, \ldots \frac{1}{30}, 0, \ldots, 0, -1]$

(g) **(5 pts)** Describe what are the model assumptions in the Gaussian Mixture Model (GMM). Is GMM a discriminative model or a generative model?

GMM is a generative model. It assumes that a data set can be categorized into a set of k Gaussian distributions, and that every sample belongs to one of the k distributions. It has no prior knowledge of the distributions and iteratively approximates an optimal distribution.