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 Mapping features into a high dimension space may allow a Perceptron to classify on non-linear or otherwise unobvious features.

(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 classifier with a higher VC dimension is more flexible and more likely to overfit the data.

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

False Independence implies conditional independence but not the other way around.

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

False SGD does not require the parameters it minimizes to reach zero, and soft-SVM can work on non-linearly separable data.

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

True This is to encourage the misclassified examples to be classified correctly in the next iteration.

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?

The model assumes F_1, F_2, and F_3 are conditionally independent given C.

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

$2 \cdot 3(3-1) = 12$

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

$2 \cdot (3^3-1) = 52$

(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.

$P(F_1 = a \mid C = +) = 1/2$
$P(F_1 = b \mid C = +) = 0$
$*P(F_1 = c \mid C = +) = 1/2$

$P(F_1 = a \mid C = -) = 1/3$
$P(F_1 = b \mid C = -) = 1/3$
$*P(F_1 = c \mid C = -) = 1/3$

$P(F_2 = a \mid C = +) = 1/2$
$P(F_2 = b \mid C = +) = 0$
$*P(F_2 = c \mid C = +) = 1/2$

$P(F_2 = a \mid C = -) = 1/3$
$P(F_2 = b \mid C = -) = 0$
$*P(F_2 = c \mid C = -) = 2/3$

$P(F_3 = a \mid C = +) = 1/2$
$P(F_3 = b \mid C = +) = 0$
$*P(F_3 = c \mid C = +) = 1/2$

$P(F_3 = a \mid C = -) = 1/3$
$P(F_3 = b \mid C = -) = 1/3$
$*P(F_3 = c \mid C = -) = 1/3$

*implied

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

Logistic regression uses continuous inputs while Naive Bayes uses discrete inputs. Also, logistic regression doesn't make the conditional independence assumption that Naive Bayes does.

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.

$$L(p_1, p_2 \mid H_1, T_1, H_2, T_2)$$

$$= P(H_1, T_1, H_2, T_2 ; p_1, p_2)$$

$$= \binom{n}{H_1, T_1, H_2} \left(\frac{1}{2}p_1\right)^{H_1} \left(\frac{1}{2}(1-p_1)\right)^{T_1} \left(\frac{1}{2}p_2\right)^{H_2} \left(\frac{1}{2}(1-p_2)\right)^{T_2}$$

(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).

$$\ln P(H_1, T_1, H_2, T_2 ; p_1, p_2) = \ln \binom{n}{H_1, T_1, H_2} + H_1 \ln \tfrac{1}{2}p_1 + T_1 \ln \tfrac{1}{2}(1-p_1) + H_2 \ln \tfrac{1}{2}p_2 + T_2 \ln \tfrac{1}{2}(1-p_2)$$

$$\frac{\partial}{\partial p_1} \ln P(H_1, T_1, H_2, T_2 ; p_1, p_2) = \frac{H_1}{p_1} + \frac{T_1}{p_1 - 1}$$

$$0 = H_1(p_1 - 1) + T_1 p_1$$

$$H_1 = (H_1 + T_1) p_1$$

$$\hat{p}_1 = \frac{H_1}{H_1 + T_1}$$

$$\frac{\partial}{\partial p_2} \ln P(H_1, T_1, H_2, T_2 ; p_1, p_2) = \frac{H_2}{p_2} + \frac{T_2}{p_2 - 1}$$

$$0 = H_2(p_2 - 1) + T_2 p_2$$

$$H_2 = (H_2 + T_2) p_2$$

$$\hat{p}_2 = \frac{H_2}{H_2 + T_2}$$

(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.

$$L(p_1, p_2, p_{white} \mid H, T)$$

$$= P(H, T; p_1, p_2, p_{white})$$

$$= \binom{n}{H} (p_{white} p_1 + (1-p_{white}) p_2)^H (p_{white}(1-p_1) + (1-p_{white})(1-p_2))^T$$

This function is challenging to maximize because it is difficult to separate the parameters p_1, p_2, and p_{white}, unlike in the last part where we were able to do so by taking a logarithm.

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

The function can be maximized using an iterative procedure:

1) Guess some $0 < p_1 < 1$, $0 < p_2 < 1$, and $0 < p_{white} < 1$.

2) Compute the likelihoods based on the observations.

3) Recompute p_1, p_2, and p_{white} as marginal probabilities (sum over joint probabilities).

4) Repeat 2&3 until parameters converge.

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* {"inner(dot) product", "sum"} in some feature space, of the feature representations that correspond to x and 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.

  (from (i))

  Let $k_1(\vec{x}, \vec{x}') = \phi(\vec{x}) \cdot \phi(\vec{x}')$ and $\phi'(\vec{x}) = \sqrt{c}\,\phi(\vec{x})$.

  $k(\vec{x}, \vec{x}') = ck_1(\vec{x}, \vec{x}') = c\phi(\vec{x}) \cdot \phi(\vec{x}') = \sqrt{c}\,\phi(\vec{x}) \cdot \sqrt{c}\,\phi(\vec{x}') \overset{\checkmark}{=} \phi'(\vec{x}) \cdot \phi'(\vec{x}')$

- **(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.

  Let $k_1(\vec{x}, \vec{x}') = \phi_1(\vec{x}) \cdot \phi_1(\vec{x}')$, $k_2(\vec{x}, \vec{x}') = \phi_2(\vec{x}) \cdot \phi_2(\vec{x}')$, $\phi'(\vec{x}) = \langle \ldots \phi_1(\vec{x}), \ldots \phi_2(\vec{x}) \rangle$

  ↑ concatenate $\phi_1(\vec{x})$ and $\phi_2(\vec{x})$

  $k(\vec{x}, \vec{x}') = k_1(\vec{x}, \vec{x}') + k_2(\vec{x}, \vec{x}')$

  $= \phi_1(\vec{x}) \cdot \phi_1(\vec{x}') + \phi_2(\vec{x}) \cdot \phi_2(\vec{x}') \overset{\checkmark}{=} \phi'(\vec{x}) \cdot \phi'(\vec{x}')$

(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).

$$\min_{w,\xi_i,b} \frac{1}{2}\vec{w}^T\vec{w} + \frac{C}{2}\sum_i \xi_i^2 + D \cdot \sum_i \text{sgn}\left[(1-\xi_i) - y_i(\vec{w}^T\vec{x_i} + b)\right]$$

D = a large number

$\text{sgn}(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0 \end{cases}$

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.

If there is a $\xi_i < 0$ that is part of an optimal solution, then another optimal solution can be obtained by replacing ξ_i with $|\xi_i|$. Doing so doesn't change the value of $\frac{1}{2}\vec{w}^T\vec{w} + \frac{C}{2}\sum_i \xi_i^2$ since ξ_i is squared, and the first constraint $y_i(\vec{w}^T\vec{x_i} + b) \geq 1 - \xi_i$ is still satisfied because $(1 - |\xi_i|) < (1 - \xi_i)$.

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.

  none, we get a trivial solution.

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

  0 and -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?

The VC dimension of H is 2. It can shatter 2 points:

But it cannot separate this labeling of 3 points:

(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 represents an error and δ represents a probability. If we have at least m examples, then with probability δ we will have no more than ϵ error. The inequality allows us to have an upper bound on the error with some probability.

(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.)

To reduce the error by a factor of $\frac{32}{15}$, the number of examples needs to be increased by the same factor.

$25 \times \frac{32}{15}$

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?

MAP allows us to incorporate knowledge of a prior distribution while MLE does not. MAP is reduced to MLE if we don't have information about the prior.

(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.)

$\ln P(y;\theta) = y \ln\theta - \theta + \ln y!$

$\frac{\partial}{\partial\theta} \ln P(y;\theta) = \frac{y}{\theta} - 1$

$0 = \frac{y}{\theta} - 1$

$1 = \frac{y}{\theta}$

$\theta = y$

(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 $\phi_\beta(\cdot)$.

$$k_\beta(\vec{x};\vec{z}) = \beta^2 + 2\beta\vec{x}\cdot\vec{z} + (\vec{x}\cdot\vec{z})^2$$

$$= \beta^2 + 2\beta\sum_i x_i z_i + \left(\sum_i x_i z_i\right)^2$$

$$= \beta^2 + 2\beta\sum_i x_i z_i + \sum_i x_i z_i \sum_j x_j z_j$$

$$= \beta\beta + \sum_i \sqrt{2\beta}\, x_i \sqrt{2\beta}\, z_i + \sum_i\sum_j x_i x_j z_i z_j$$

$$= \left\langle \beta, \underbrace{\sqrt{2\beta}\,x_1, \ldots, \sqrt{2\beta}\,x_i, \ldots}_{\forall i}, \underbrace{x_1x_1, \ldots x_{ij}, \ldots}_{\forall i_{ij}} \right\rangle \cdot \left\langle \beta, \underbrace{\sqrt{2\beta}\,z_1, \ldots, \sqrt{2\beta}\,z_i, \ldots}_{\forall i}, \underbrace{z_1z_1, \ldots, z_iz_j, \ldots}_{\forall i_{ij}} \right\rangle$$

$$\phi_\beta(\vec{x}) = \left\langle \beta, \underbrace{\sqrt{2\beta}\,x_1, \ldots, \sqrt{2\beta}\,x_i, \ldots}_{\forall i}, \underbrace{x_1x_1, \ldots, x_{ij}, \ldots}_{\forall i_{ij}} \right\rangle.$$

In $\mathbb{R}^3$: $\phi_\beta(\vec{x}) = \langle \beta, \sqrt{2\beta}\,x_1, \sqrt{2\beta}\,x_2, \sqrt{2\beta}\,x_3, x_1^2, x_1x_2, x_1x_3, x_2x_1, x_2^2, x_2x_3, x_3x_1, x_3x_2, x_3^2 \rangle$

equivalent to: $\langle \beta, \sqrt{2\beta}\,x_1, \sqrt{2\beta}\,x_2, \sqrt{2\beta}\,x_3, x_1^2, \sqrt{2}\,x_1x_2, \sqrt{2}\,x_1x_3, x_2^2, \sqrt{2}\,x_2x_3, x_3^2 \rangle$

(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?

I recommend a generative classifier. Discriminative classifiers don't learn the joint probabilities needed for density estimation.

(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 $\{0,1\}^{100}$.

Represent true as 1 and false as 0.

$h(\vec{x}) = \sum_{i=1}^{30} x_i - 9.5$

(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 which assumes that the data are centered around several means, and each mean is that of a Gaussian distribution.