Student Name: ____________________

Student Number: ____________________

School: ____________________

Year of Entrance: ____________________

ShanghaiTech University Final Examination Cover Sheet

Academic Year: 2020 to 2021 Term: Spring

Course-offering School: School of Information Science and Technology

Instructor: Lu Sun

Course Name: Optimization and Machine Learning

Course Number: SI151

Exam Instructions for Students:

1. All examination rules must be strictly observed throughout the entire test, and any form of cheating is prohibited.

2. Other than allowable materials, students taking closed-book tests must place their books, notes, tablets and any other electronic devices in places designated by the examiners.

3. Students taking open-book tests may use allowable materials authorized by the examiners. They must complete the exam independently without discussion with each other or exchange of materials.

For Marker's Use:

Section	I	II	III	IV	V	VI	VII	VIII	IX	Total
Marks										
Recheck										

Marker's Signature:
Date:

Rechecker's Signature:
Date:

I Basics [20 points]

Note: in the following questions, you may mark one or more than one of the choices.

1. **[2 points]** Linear regression estimator has the smallest variance among all unbiased estimators.

   (a) True

   (b) False

2. **[2 points]** Since classification is a special case of regression, logistic regression is a special case of linear regression.

   (a) True

   (b) False

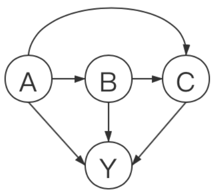
3. **[2 points]** The training error of 1-nearest neighbor classifier is 0.

   (a) True

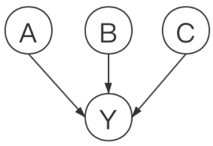
   (b) False

4. **[2 points]** Suppose that you have a dataset with 3 categorical input attributes A, B and C. There is one categorical output attribute Y. You are trying to learn a Naive Bayes Classifier for predicting Y. Which of these Bayes Net diagrams represent(s) the naive bayes classifier assumption?

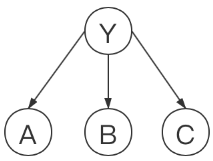
   (a)



   (b)



   (c)



   (d)

5. **[2 points]** In each round of AdaBoost, the misclassification penalty for a particular training observation is increased going from round t to round $t + 1$ if the observation was:

   (a) classified incorrectly by the weak learner trained in round t.

   (b) classified incorrectly by the full ensemble trained up to round t.

   (c) classified incorrectly by a majority of the weak learners trained up to round t.

6. **[2 points]** AdaBoost minimizes an exponential loss function.

   (a) True

(b) False

7. [**2 points**] What statement(s) are true about the expectation-maximization (EM) algorithm?

   (a) It requires some assumption about the latent probability distribution.

   (b) Comparing to a gradient descent algorithm that optimizes the same objective function as EM, EM may only find a local optima whereas the gradient descent will always find the global optima.

   (c) The EM algorithm maximizes a lower bound of the marginal likelihood $P(\mathcal{D}; \boldsymbol{\theta})$

   (d) The algorithm assumes some that some of the data generated by the probability distribution is not observed.

8. [**2 points**] The SVM learning algorithm is guaranteed to find the globally optimal hypothesis with respect to its object function.

   (a) True

   (b) False

9. [**2 points**] Which statement(s) are true about the K-means algorithm?

   (a) It is a clustering algorithm.

   (b) It is an EM algorithm.

   (c) It assumes the data is from a mixture of Gaussian distributions.

   (d) It is a soft EM algorithm, where all possible hidden attributes are considered in the E step.

   (e) It is guaranteed to converge to the global optimum.

   (f) It is a convex optimization problem.

10. [**2 points**] Query strategy plays a key role in active learning. Generally, the following query strategies can be selected: uncertainty sampling, query-by-committee, expected model change, expected error reduction, variance reduction, density-weighted methods. Which of the following option(s) is(are) reasonable method(s) of query strategies?

   (a) Least confident method, which is to select samples that have a low maximum classification probability.

   (b) Margin sampling method, which is to select samples of data that can easily be classified into two categories, or that have a similar probability of being classified into two categories.

   (c) Entropy method, which is to select samples of data that have high entropy in a particular system. (The definition of entropy is $-\sum_i P_\theta(y_i \mid x) \cdot \ln P_\theta(y_i \mid x)$.)

   (d) Expected loss method, which is to select samples of data that will cause the loss function to reduce the least by adding a sample.

   Your answer:

II Regression and Probability Estimation [12 points]

Consider real-valued variables X and Y, in which Y is generated conditional on X according to

$$Y = aX + \epsilon, \text{ where } \epsilon \sim \mathcal{N}(0, \sigma^2).$$

Here ϵ is an independent variable, called a noise term, which is drawn from a Gaussian distribution with mean 0, and variance σ^2. This is a single variable linear regression model, where a is the only weight parameter. The conditional probability of Y has distribution $p(Y|X, a) \sim \mathcal{N}(aX, \sigma^2)$, so it can be written as:

$$p(Y|X, a) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(Y - aX)^2\right).$$

The following questions are all about this model.

1. **[4 points]** Assume we have a training dataset of n pairs (X_i, Y_i), $i = 1, 2, ..., n$. Which one(s) of the following equations correctly represent(s) the Maximum Likelihood Estimation (MLE) problem for estimating a? (You may mark one or more than one of the choices.)

   (a) $\arg\max_a \sum_i \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\right)$

   (b) $\arg\max_a \prod_i \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\right)$

   (c) $\arg\max_a \sum_i \exp\left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\right)$

   (d) $\arg\max_a \prod_i \exp\left(-\frac{1}{2\sigma^2}(Y_i - aX_i)^2\right)$

   (e) $\arg\max_a \frac{1}{2}\sum_i (Y_i - aX_i)^2$

   (f) $\arg\min_a \frac{1}{2}\sum_i (Y_i - aX_i)^2$

2. **[4 points]** Derive the maximum likelihood estimate of the parameter a in terms of the training data (X_i, Y_i), $i = 1, 2, ..., n$. You are recommended to start with the simplest form of the problem you found above.

3. **[4 points]** Let's put a prior on a, for example, $a \sim \mathcal{N}(0, \lambda^2)$, i.e.,

$$p(a|\lambda) = \frac{1}{\sqrt{2\pi}\lambda} \exp(-\frac{1}{2\lambda^2}a^2).$$

   (a) Under which case(s) that the estimated value with MLE and Maximum A Posterior (MAP) will become closer, in other words, $|a^{MLE} - a^{MAP}|$ will decrease? (You may mark one or more than one of the choices.)

      i. As $\lambda \to \infty$

      ii. As $\lambda \to 0$

      iii. Fix λ and as number of training samples $n \to \infty$

   (b) Assume $\sigma = 1$, and a fixed prior parameter λ. Solve for the MAP estimate of a:

$$\arg\max_a \left[\log p(Y_1, ..., Y_n|X_1, ..., X_n, a) + \log p(a|\lambda)\right].$$

   Your solution should be in terms of X_i's Y_i's and λ.

Your answer:

III Linear Classification [10 points]

Given the input continuous variable X and the output categorical variable Y, suppose that:

- We know $P(Y = k) = \pi_k$ exactly.
- $P(X = \mathbf{x} \mid Y = k)$ is multivariate normal distribution with density:

$$f_k(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu_k)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_k)}, \quad \mathbf{x} \in \mathbb{R}^p,$$

  where μ_k is the mean of the inputs for category k and $\mathbf{\Sigma}$ is the covariance matrix.

Answer the questions below:

1. **[3 points]** What is the Bayes classifier (maximize the probability of category k, given the input $\mathbf{x}$)?

2. **[3 points]** Please derive the linear discriminant function $\delta_k(\mathbf{x})$, and explain how to predict the category of input $\mathbf{x}$.

3. **[4 points]** Show what is the decision boundary between category k and l given the input $\mathbf{x}$. For some vectors $\mathbf{w}$ and scalar b, the decision boundary can be expressed as $\mathbf{w}^T\mathbf{x} + b = 0$. Find the entries of the vector $\mathbf{w}$ and the value of b in terms of class priors and parameters.

Your answer:

IV Graphical Model [**10 points**]

Consider the following Bayesian Network, in which all variables are boolean.

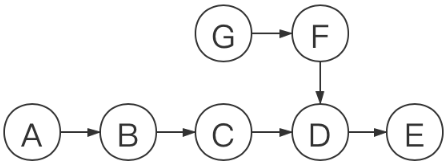


Figure 1: Bayesian network with seven boolean variables.

1. [**4 points**] Write the expression for the joint likelihood of the network in its factorized form.

2. [**3 points**] Let $X = \{C\}, Y = \{B, D\}, Z = \{A, E, F, G\}$. Is $X \perp Z | Y$?, If yes, explain why. If no, show a path from X to Z is not blocked.

3. [**3 points**] Directly prove that $A \perp C | B$ without using D-separation.

Your answer:

V Kernel Methods [8 points]

Kernel functions implicitly define some mapping function $\phi(\cdot)$ that transforms an input instance $x \in \mathbb{R}^d$ to a high dimensional feature space Q, by giving the form of dot product in $Q : K(x_i, x_j) = \phi(x_i)\dot{\phi}(x_j)$. Assume we use radial basis kernel function

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{1}{2}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right).$$

1. **[4 points]** Prove that for arbitrary two input instances x_i and x_j, the squared Euclidean distance of their corresponding points in the feature space Q is less than 2, i.e.,

$$\|\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j)\|^2 < 2.$$

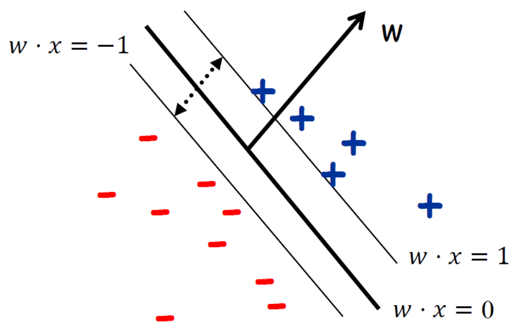
2. **[2 points]** The dimensionality of the feature map generated by radial basis kernel is infinity.
   (a) True
   (b) False

3. **[2 points]** The dimensionality of the feature map generated by polynomial kernel (e.g., $K(x, y) = (1 + xy)^d$) is polynomial w.r.t. the power d of the polynomial kernel.
   (a) True
   (b) False

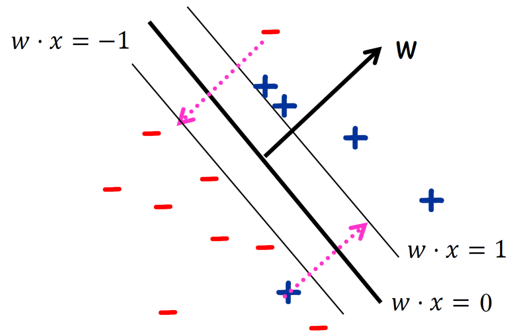
Your answer:

VI Support Vector Machines [10 points]

Support vector machines (SVM) are supervised learning models, that directly optimize for the maximum margin separator. Fig. 2 shows an example of maximum margin separator over a dataset $S = \{(x_i, y_i)\}_{i=1}^{n}$, in which $x_i \in \mathbb{R}^2$ and $y_i \in \{-1, 1\}$ denote the i-th sample and the i-th label ($\forall i$), respectively, in both separable case (Fig.2(a)) and non-separable case (Fig.2(b)). For simplicity, here we assume that the dataset S has been standardized, and thus the bias can be omitted in the linear model. In Fig. 2, "+" and "-" denote the samples with labels "1" and "-1", respectively, and $\mathbf{w}$ is the normal vector of the maximum margin separator $\mathbf{w}^\top x = 0$. You need to derive the linear optimization problem of SVM in both separable case and non-separable case.
Note: correctly giving the results without detailed derivation will get half the points.



(a) Separable case



(b) Non-separable case

Figure 2: Maximum margin separator.

1. [**3 points**] Derive the constraint optimization problem of SVM in the separable case shown in Fig. 2(a).

2. [**3 points**] Extend the results in (a) to handle the non-separable case shown in Fig. 2(b).

3. [**4 points**] Show the unconstraint form of the above problem and determine the convexity. You need to explain the reason for your answer.

Your answer:

VII Principal Component Analysis [**9 points**]

Given 3 data points in 2D space: $(1, 1)$, $(2, 2)$, $(3, 3)$, answer the following questions:

1. [**3 points**] What is the first principle component?

2. [**3 points**] If we want to project the original data points into 1D space by principle component you choose, what is the variance of the projected data?

3. [**3 points**] For the projected data in 1D space, now if we represent them in the original 2D space, what is the reconstruction error?

Your answer:

VIII Neural Networks [9 points]

Consider the network shown in the figure. All of the hidden units use the rectified linear unit (ReLU): $h_i = \max(z_i, 0)$. We are trying to minimize a cost function C which depends only on the activation of the output unit y. The unit h_1 (marked with $\star$) receives an input of -1 on a particular training case, so its output is 0. Based only on this information, which of the following weight derivatives are guaranteed to be 0 for this training case? Write TRUE or FALSE for each. (Hint: don't work through the backpropagation computations, instead, think about what do the partial derivatives really mean.)
Note: correct answers without explanation will get half the points.

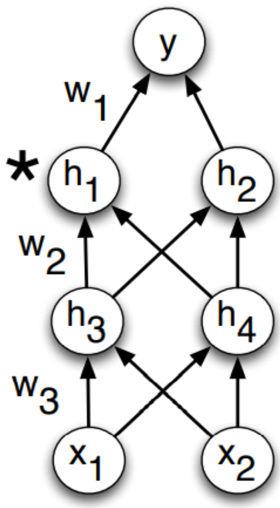


Figure 3: Neural Network with four layers. (Note: Each of $w1$, $w2$, and $w3$ refers to the weight on a single connection, not the whole layer.)

1. **[3 points]** $\partial C/\partial w_1 = 0$: ____, your explanation:

2. **[3 points]** $\partial C/\partial w_2 = 0$: ____, your explanation:

3. **[3 points]** $\partial C/\partial w_3 = 0$: ____, your explanation:

Your answer:

IX Convex Sets and Convex Functions [12 points]

In this problem, you should first write down whether the set or the function is convex or non-convex, then either prove the set or the function is convex or provide an example to show that it's non-convex.
Note: correct answers without proof will get half the points.

1. [**6 points**] Determine the convexity of the following sets:

   (a) Polyhedra:
   $$\{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} \preceq \mathbf{b}, \mathbf{C}\mathbf{x} = \mathbf{d}\},$$
   where $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{C} \in \mathbb{R}^{p \times n}$, and $\mathbf{d} \in \mathbb{R}^p$.

   (b) Positive semidefinite cone:
   $$\mathbb{S}^n_+ = \{\mathbf{X} \in \mathbb{S}^n \mid \mathbf{X} \succeq 0\},$$
   where $\mathbb{S}^n$ denotes the set of symmetric matrices in $\mathbb{R}^{n \times n}$. Here $\mathbf{X} \succeq 0$ represents the generalized inequality on matrices, indicating $\mathbf{z}^\top \mathbf{X} \mathbf{z} \geq 0$, $\forall \mathbf{z} \in \mathbb{R}^n$.

2. [**6 points**] Determine the convexity of the following functions:

   (a) Lasso objective:
   $$f(\mathbf{x}) = ||\mathbf{A}\mathbf{x} - \mathbf{b}||^2 + \lambda ||\mathbf{x}||_1,$$
   where $\mathbf{A} \in \mathbb{R}^{m \times n}$ ($\mathbf{A}^\top \mathbf{A} \in \mathbb{S}^n_+$), $\mathbf{b} \in \mathbb{R}^m$, $\lambda > 0$, and $|| \cdot ||$ and $|| \cdot ||_1$ denote ℓ_2-norm and ℓ_1-norm, respectively.

   (b) Weighted log barrier for linear inequalities:
   $$f(\mathbf{x}) = -\sum_{i=1}^{m} c_i \log(b_i - \mathbf{a}_i^\top \mathbf{x}),$$
   with $\mathbf{dom} f = \{\mathbf{x} \mid \mathbf{a}_i^\top \mathbf{x} < b_i,\ i = 1, 2, ..., m\}$. Here $\mathbf{a}_i, \mathbf{x} \in \mathbb{R}^n$, and $c_i > 0$ denotes the weighting coefficient, $i = 1, 2, ..., m$.

Your answer: