DS-GA-1003: Machine Learning (Spring 2021)

Midterm Exam (March 23 – March 24)

- You should finish the exam within **2 hours** once it is started and submit on Gradescope by **5:20pm EST on March 24**.
- You can refer to textbooks, lecture slides, and notes. However, searching answers online and collaboration are not allowed.
- Please write your solution on a separate sheet either by hand or by typing. The final submission must be in PDF format and each question (six in total) should start on a new page.
- Using the provided template is optional. You can find the template here: https://tinyurl.com/3fhjtxwj

Question	Points	Score
Generalization	15	
MLE	15	
Optimization	15	
Regularization	15	
SVM	15	
Kernels	15	
Total:	90	

1. **Generalization and risk decomposition.** Alice and Bob are trying to build a classifier to predict whether a student will take DS1003 next spring. They have collected data of students who did and did not take DS1003 in the past.

   (a) (2 points) They first split the data into train, validation, and test sets. In 1–2 sentences, explain the difference between the test set and the validation set.

   Solution: The validation set is used for model selection and the test set is used for final evaluation.

   (b) (2 points) After training a linear classifier, they find that both the training error and the validation error are very high. Alice suggests that they try to reduce training error first. To decrease training error, they would want to use

   A. more training data

   B. less training data

   (c) (2 points) After changing the amount of data, the training error is still quite high. Bob figures that there might be something wrong with the features. To reduce training error, they should use

   A. more features

   B. fewer features

   (d) (3 points) After some feature engineering, the training error finally goes down. Encouraged by this result, Bob suggests to use student NetID as a feature. Is this a good idea? Why or why not?

   Solution: Using NetID, the model can achieve zero training error by memorizing the label of each student, so it will cause overfitting.

   (e) (1 point) When checking the validation error, they find that it is much higher than the training error. This is a sign of **overfitting / poor generalization**.

   (f) (3 points) To reduce the gap between training and validation error, Alice makes the following suggestions. Which one(s) would you agree with if you were Bob? (Select all correct answers)

   A. Collect more training data

   B. Add more features

   C. Add regularization

   D. Use a non-linear classifier

   (g) (2 points) Finally, Alice and Bob are happy with their classifier and send it to the department head with a reported test accuracy of 90%. If you were the department head, do you expect this classifier to have similar prediction accuracy next spring? Why or why not?

Solution: The accuracy is expected to be lower as the cohort next spring will be different from past students. A positive answer is also considered correct if they assume that students won't change too much within one year.

2. **MLE and probabilistic models.** Alice and Bob are still trying to predict who will take DS1003 next spring as in question 1. This time however they would like to estimate the probability of a student to enroll in DS1003. At first, they consider the students to be independent and identical and collect N outcomes $\mathcal{D} = y_i{}_{i=1}^N$ with $y_i = 1$ if student i took DS1003 and $y_i = 0$ if not.

   (a) (2 points) What is an appropriate parametric family of distributions to model the desired probability.

   A. Gaussian family

   B. Beta family

   C. Bernoulli family

   D. Poisson family

   (b) (1 point) We call $\theta \in \Theta$ the parameter of the distribution in the family noted $\{p(y|\theta), \theta \in \Theta\}$. What is the appropriate Θ?

   Solution: $\Theta = [0, 1]$.

   (c) (3 points) Alice and Bob are going to use Maximum Likelihood Estimation to chose a value for θ. What is the likelihood of $\mathcal{D}$ given θ as a function of the y_i and θ?

   Solution: $p(\mathcal{D}|\theta) = \prod_{i=1}^N \theta^{y_i}(1-\theta)^{1-y_i}$.

   (d) (1 point) At this point, Alice and Bob think of differentiating between the students using features they collected. We denote by $x \in \mathbf{R}^d$ the features and the training data set is now $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$. They choose a conditional family with a linear predictors parametrized by $w \in \mathbf{R}^d$ such that under their model $p(y = 1|x, w) = f(w^T x)$. The transfer funcion f must be a mapping from which space to which space?

   A. $[0, 1] \to \mathbf{R}$

   B. $[0, 1] \to [0, 1]$

   C. $\mathbf{R} \to [0, 1]$

   (e) (3 points) They choose $f(x) = 1/(1 + e^{-x})$. Write down the log-likelihood of data set $\mathcal{D}$ as a function of w, x_is and y_is.

Solution: $p(\mathcal{D}|w) = \prod_{i=1}^{N} f(w^T x_i)^{y_i}(1 - f(w^T x_i))^{1-y_i}$
$\log p(\mathcal{D}|w) = -\sum_{i=1}^{N} y_i \log(1 + e^{-w^T x_i}) + (1 - y_i)\log(1 + e^{w^T x_i})$.

(f) (3 points) What is the optimization problem to solve to get a $\hat{w}$ by maximum likelihood estimation and which method would you use to solve it? (Write the optimization problem in terms of $\hat{w} = \arg\max_w \ldots$.)

Solution: $\hat{w} = \arg\max_w - \sum_{i=1}^{N} y_i \log(1 + e^{-w^T x_i}) + (1 - y_i)\log(1 + e^{w^T x_i})$
There is no closed form a prior, we can use gradient descent optimization.

(g) (2 points) Alice and Bob have access to a test and validation set on top of the training data they used so far. They decide to finally test different possible transfer functions f_1, f_2 and f_3 and obtain by maximum likelihood corresponding $\hat{w}_1, \hat{w}_2$ and $\hat{w}_3$. They compute the likelihood of the different sets for the different models they obtained and collect the result in an array:

	training set	validation set	testing set
$f_1,\ \hat{w}_1$	0.55	0.45	0.43
$f_2,\ \hat{w}_2$	0.57	0.43	0.43
$f_3,\ \hat{w}_3$	0.56	0.42	0.45

Which model would you choose and why?

Solution: model 1 since it has the maximum likelihood on the validation set.

3. **Loss functions and optimization.** Your boss asks you to build a classifier to detect fake news on COVID-19. You decide to use a linear classifier for simplicity:

$$f(x) = \begin{cases} 1 & \text{if } w \cdot \varphi(x) \geq 0 \\ -1 & \text{otherwise} \end{cases},$$

where φ is the feature map and $w \in \mathbb{R}^d$ is the parameter. We denote the label by $y \in \{-1, +1\}$. A positive prediction (+1) means x is fake news.

(a) You will use the familiar ERM objective and SGD to learn the parameter w. For each of the following loss functions, explain whether it is suitable for the classification task and the learning framework. If yes, write down the (sub)gradient for a single example.

i. (2 points)

$$\ell(x, y, w) = \mathbb{I}\left[(\varphi(x) \cdot w)y \leq 0\right],$$

where $\mathbb{I}(\cdot)$ is the indicator function.

Solution: No. (Sub)gradient is zero everywhere.

ii. (2 points)

$$\ell(x, y, w) = \min\left((\varphi(x) \cdot w)y, 0\right)$$

Solution: No. Loss decreases when prediction is more incorrect (small margin).

iii. (2 points)

$$\ell(x, y, w) = 1 - (\varphi(x) \cdot w)y$$

Solution: No. Optimal w is either zero or does not exist.

iv. (2 points)

$$\ell(x, y, w) = \begin{cases} 1 - 2(\varphi(x) \cdot w)y & \text{if } (\varphi(x) \cdot w)y \leq 0 \\ (1 - (\varphi(x) \cdot w)y)^2 & \text{if } 0 \leq (\varphi(x) \cdot w)y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Solution: Yes.

$$\nabla_w \ell(x, y, w) = \begin{cases} -2y\varphi(x) & \text{if } (\varphi(x) \cdot w)y \leq 0 \\ -2y(1 - (\varphi(x) \cdot w)y)\varphi(x) & \text{if } 0 \leq (\varphi(x) \cdot w)y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

(b) (2 points) You recall from DS1003 that hinge loss is commonly used for classification tasks:

$$\ell(x, y, w) = \max(1 - (\varphi(x) \cdot w)y, 0),$$

and decide to give it a try. After several SGD epochs, you find that the average hinge loss of the training examples is 0.2. Your boss does not know hinge loss though and asks about the accuracy. What can you say about the training accuracy? Your answer should be one sentence.

Solution: The training accuracy is at least 0.8 because hinge loss is an upper bound of the 0-1 loss.

(c) (2 points) You soon realized that there are ten times more real news (negative examples) than fake news (positive examples) in your dataset because it is expensive to have fake news annotated. In this case, is a high accuracy meaningful? Why or why not?

Solution: No. It is easy to get high accuracy (e.g. 90%) by predicting all examples as real news.

(d) (3 points) Your next job is to design a new loss function such the loss is zero if the margin $(y\varphi(x) \cdot w)$ is larger than 1, and the loss of a positive example is 10 times the loss of a negative example given the same margin because you do not want to miss any fake news! Write down your loss function $\ell(x, y, w)$ and its gradient $\nabla_w \ell(x, y, w)$.

Solution:

$$\ell(x, y, w) = \begin{cases} 10 \max(1 - y\varphi(x) \cdot w, 0) & \text{if } y = 1 \\ \max(1 - y\varphi(x) \cdot w, 0) & \text{if } y = -1 \end{cases}$$

Solution:

$$\nabla_w \ell(x, y, w) = \begin{cases} 0 & \text{if } y\varphi(x) \cdot w - 1 \geq 0 \\ -10\varphi(x) & \text{if } y = 1 \text{ and } y\varphi(x) \cdot w - 1 < 0 \\ \varphi(x) & \text{if } y = -1 \text{ and } y\varphi(x) \cdot w - 1 < 0 \end{cases}$$

4. **Regularization.** Recall that if we consider the perceptron algorithm as SGD using a certain loss function, we can define the perceptron loss:

$$\ell(x, y, w) = \max(-yw^T x, 0) \ ,$$

and the hypothesis space is given by

$$\mathcal{H} = \left\{ f \mid f(x) = w^T x, w \in \mathbb{R}^d \right\} \ .$$

(a) (3 points) Given linearly separable data, does the perceptron algorithm find a unique solution w^*? Why or why not?

Solution: No. The solution depends on the order of the examples.

(b) (3 points) Now consider minimizing the ERM objective using the perceptron loss with L2 regularization:

$$J(w) = \sum_{i=1}^{n} \max(-y_i w^T x_i, 0) + c\|w\|_2^2$$

Is the solution unique?

Solution: Yes.

(c) (3 points) Give a solution w that achieves the minimum $J(w)$ but cannot separate the training data.

Solution: The zero vector.

(d) (3 points) To avoid the degenerate solution in the previous question, let's add a constraint $\|w\|_2^2 \geq 1$. Write the Lagrangian for the problem of minimizing $J(w)$ subject to the constraint and specify which variable(s) is the lagrangian multiplier(s). (You can use $J(w)$ in the expression)

Solution:

$$L(w, \lambda) = J(w) + \lambda(1 - \|w\|_2^2)$$

(e) (3 points) Is the new constraint minimization problem convex? Why or why not? Recall that convex problems must have convex objective and convex feasible set.

Solution: No. $\{w \mid \|w\|_2^2 \geq 1\}$ is not a convex set.

5. **Support Vector Machines.**

(a) (3 points) Recall that the SVM objective is equivalent to minimizing the average hinge loss with L2 regularization. What if we minimize the hinge loss without regularization? Explain why the solution is not unique without regularization. (You can assume the data is linearly separable.)

Solution: Given w^* that achieves zero training error, we can arbitrarily scale it without changing the total loss.

(b) Recall the hard-margin SVM objective:

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|_2^2 \\ \text{s.t.} \quad & y_i(w^T x_i + b) \geq 1 \ \forall i \end{aligned}$$

The constraints says that the (functional) margin of each example is at least 1. If we change the constraint to require the margin to be at least c ($c > 0$), i.e. solving

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|_2^2 \\ \text{s.t.} \quad & y_i(w^T x_i + b) \geq c \; \forall i \end{aligned}$$

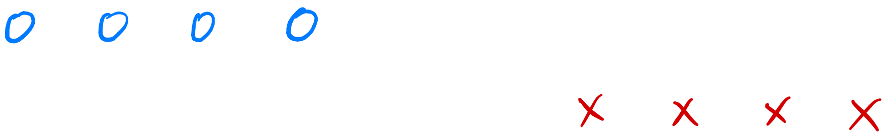
i. (3 points) Would it change the separating hyperplane? Why or why not?

Solution: No. It would only scale w and b.

ii. (3 points) Let w^* be the solution of the original hard-margin SVM, and w' be the solution of the modified problem with margin at least c. Write an expression of w' using w^*.

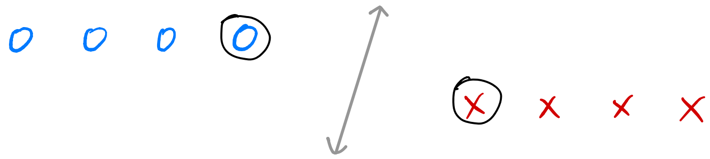
Solution: $w' = cw^*$

(c) Consider the following dataset:



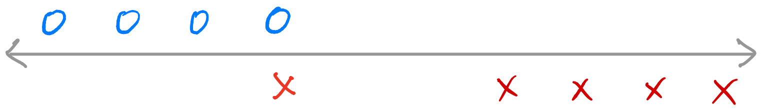
where circles are positive examples and crosses are negative examples.

i. (2 points) Draw the decision boundary given by a linear (hard-margin) SVM trained on this dataset, and circle the support vectors.



Solution:

ii. (2 points) Now add a single example to the training set such that all examples would be support vectors.

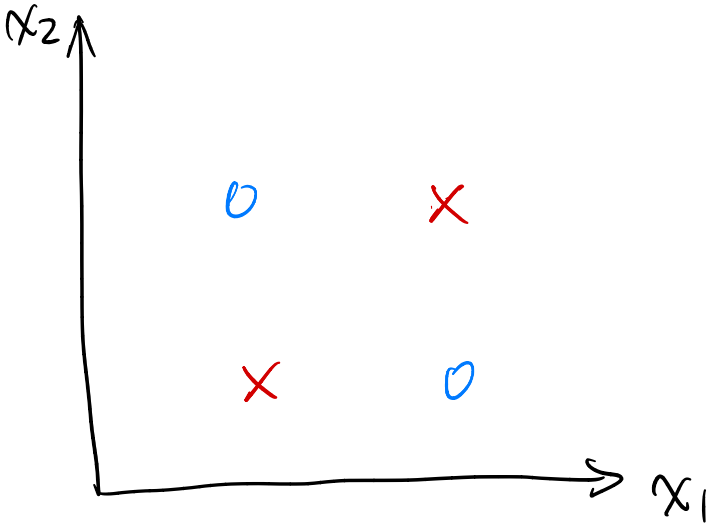


Solution:

(d) (2 points) Given your observation in the previous question, explain why soft-margin SVM may be preferred even if the data is linearly separable.

Solution: Robust to outliers.

6. **Kernels.** Consider the following dataset:



Each point is an example in $\mathbb{R}^2$, i.e. $x = (x_1, x_2)$. Circles are positive examples and crosses are negative examples.

(a) (3 points) What's the maximum training accuracy we can achieve on this dataset if we use a linear classifier?

Solution: 75%

(b) (3 points) Add a single feature to make the dataset linearly separable.

Solution: There are many answers, e.g. $(x_1 - x_2)^2$.

(c) (3 points) Suppose we are going to use SVM to solve the binary classification problem. Which of the following kernel(s) can separate the data? (Select all correct answers)

A. linear kernel

B. polynomial kernel with degree 3

C. Gaussian kernel

(d) Now consider the general setting where $x \in \mathbb{R}^p$ and we have n training examples. Suppose we learn a SVM with quadratic kernel and find that there are m support vectors.

i. (1 point) At inference time, what is the computation cost to make a single prediction using the primal solution? Express your answer in m, n, p using the big-O notation.

Solution: $O(p^2)$

ii. (1 point) What if we use the dual solution (i.e. use the kernel trick)?

Solution: $O(mp)$

iii. (1 point) Given your above answers, when do we want to use the dual solution for prediction?

Solution: $m < p$

(e) (3 points) For which of the following model(s)/algorithm(s) can we use the kernel trick? (Select all correct answers)

A. Perceptron

B. Logistic regression with L2 regularization

C. Lasso regression

D. k**-nearest neighbor**

Congratulations! You have reached the end of the exam.