CS4780 Final

Spring 2018

NAME:	
Net ID:	
Email:	

1 [??] General Machine Learning

Please identify if these statements are either True or False. Please justify your answer **if false**. Correct "True" questions yield 1 point. Correct "False" questions yield two points, one for the answer and one for the justification.

1. (**T/F**)If we have a validation set, we do not need to do k-fold cross validation for hyperparameter tuning.

2. (**T/F**)As $n \to \infty$, the 1-NN error is less than thrice the error of Bayes Optimal Classifier.

3. (**T/F**) For perceptron algorithms, the order of datapoints do affect the number of misclassifications of each points.

4. (**T/F**) In MLE, the parameter we want to learn is a random variable.

5. (**T/F**) A "True Bayesian" approach of learning will learn a point estimate of the model parameter.

6. (**T/F**) Logistic Regression is a special case of Naive Bayes classifier.

7. (**T/F**) For Logistic Regression, Newton's method will return the global optimum.

8. (**T/F**) For AdaGrad, each feature has its own learning rate.

9. (**T/F**) Linear regression will perform poorly if the relationship between the label and the features is not linear .

10. (**T/F**) In a linearly separable dataset, both linear SVM with hard and soft constraints will return the same solution.

11. (**T/F**) l1 regularizer encourage sparse solutions.

12. (**T/F**) For SVM, setting the bias term to zero will increase the variance of the model

13. (**T/F**) Linear regression is kernelizable.

14. (**T/F**) The marginal likelihood of Gaussian Process is not Gaussian.

15. (**T/F**) Gaussian Process is not a parametric model .

16. (**T/F**)To use CART models, we have to scale every dimension of the your data to the range $[0, 1]$.

17. (**T/F**) In bagging, we use independent and identically distributed datasets to train our ensemble.

18. (**T/F**) For boosting, we can learn multiple base classifiers in parallel.

19. (**T/F**) For AdaBoost, the labels has to be $\{-1, +1\}$.

20. (**T/F**) Deep learning allows us to implicitly learn a fixed high dimensional feature mapping

2 [21] Bias Variance / Model Selection

1. Hyperparameter tuning is crucial in creating a powerful model. State two methods that we can use to do hyperparameter tuning for any models.

2. Suppose we are doing Gaussian Process regression. Unlike most of the models, there is a special method we can use to tune the model's hyperparameters. Explain how we could tune the hyperparameters of the Gaussian Process regression model.

3. Suppose we train a SVM classifier on n training data. After training, we obtain k support vectors and the training error (0-1 classification error) is zero.

   (a) Suppose we remove one non-support vector from the training set and train another SVM on the remaining $n-1$ training data. Then we test our new model on the removed point. What will the error be when we apply the SVM to the removed point?

   (b) Suppose we remove one support vector from the training set, train another SVM model and test the new SVM model on the removed support vector. Will the error be similar to (a)? Why or why not?

   (c) Suppose we perform Leave-One-Out Cross Validation on the training set and produce an estimate of the real test loss. Give an upper bound of the test loss in terms of n and k.

(d) Does SVM generalize better when we increase the number of training points?

(e) Does SVM generalize better when we have a large number of support vectors?

4. (13) For each of the following scenarios, determine if the model has low/high bias and variance. Explain your choice.

   (a) SVM with linear and RBF kernel.

   (b) Gaussian Naive Bayes, logistic regression

3 [21] Kernel Methods

1. Being Cornellians, your friends heard that the perceptron algorithm was invented at Cornell. They want to try classifying the following dataset using the perceptron algorithm.

Order	$\mathcal{X}$	$\mathcal{Y}$
1	$[-1, 0]^T$	$+1$
2	$[0, 0]^T$	-1
3	$[1, 0]^T$	$+1$

   (a) Your friends tried running the perceptron algorithm multiple times and the algorithm did not converge. They asked you for help. Explain to them why the algorithm would not converge on this dataset.

   (b) Your friends were convinced that the linear perceptron algorithm would not converge on this dataset and were very disappointed. However, equipped with the knowledge you learn in CS4780/5780, you know you can kernelize the perceptron algorithm to make it a more powerful algorithm. Suppose the kernel function is k. Fill in the missing pieces of the kernelized perceptron algorithm below:

Algorithm 1: Kernelized Perceptron

```
Initialize α⃗ = 0⃗ ;
while TRUE do
    m = 0 ;
    for (x_i, y_i) ∈ D do
        if _____ then
            _____;
            _____;
        end
    end
    if m = 0 then
        break
    end
end
```

(c) After looking at the dataset, you suggest your friends to use the kernelized perceptron using the quadratic kernel

$$k(\vec{x}_1, \vec{x}_2) = (\vec{x}_1^T \vec{x}_2 + 1)^2$$

Your friends run the kernelized perceptron and get $\vec{\alpha} = [1, 2, 1]^T$. However, when they applied the model to the training set, they could not get 100% accuracy. Again, knowing that you are the pro in ML, they ask for your help again. After talking to them, you realize that they might not have run the algorithm till convergence. So, you ran the kernelized perceptron algorithm with $\vec{\alpha}$ initialized to $[1, 2, 1]^T$. What is the α you get after you run the algorithm till convergence?

4 [21] CART

1. Suppose a classification tree with depth 1 can correctly classify our data. Is this classification tree a linear model? Explain your reasoning.

2. CART models are highly flexible. Suggest three ways to reduce its variance.

3. Under what condition will CART models be non-parametric.

4. Consider the following dataset: If we grow a classification tree using ID3 algorithm and Gini Impurity, what is the depth of the final tree?

Feature			Label
Go to Class	Do Projects and HWs	Pass Final Exam	Will Pass CS4780
Always	Yes	No	Yes
Always	No	Yes	Yes
Always	No	No	Yes
Occasionally	No	Yes	No
Occasionally	No	No	No
Occasionally	No	Yes	No

5 [x points] AdaBoost

In this question, we are going to explore AdaBoost using classification tree of **depth 1** as our base classifier. For your reference, we have included the following pseudo-code for AdaBoost:

Algorithm 2: AdaBoost (Assume n training points.)

Initialize $H_0 = 0; \forall i : w_i = \frac{1}{n}$;
for *t = 0: T - 1* **do**
  h = the base classifier that minimizes the weighted classification error ϵ_t :

$$\epsilon = \sum_{i:h(x_i) \neq y_i} w_i$$

  if $\epsilon < \frac{1}{2}$ **then**
    $\alpha = \frac{1}{2} \ln(\frac{1-\epsilon}{\epsilon})$;
    $H_{t+1} = H_t + \alpha h$;
    $\forall i : w_i \propto w_i e^{-\alpha h(x_i) y_i}$;
  else
    return H_t
  end
  return H_T
end

Consider the dataset in figure TODO

1. (x points) In the figure above, draw the decision boundary of first classification tree that AdaBoost would choose. Please indicate the positive and negative side of the decision boundary. (Note: Throughout the whole question, please assume that your tree would split at integer values for each branch.)

2. (x points) What is the weighted classification error **before** weights are renormalized.

3. (x points) What is the weight α of the first classification tree?

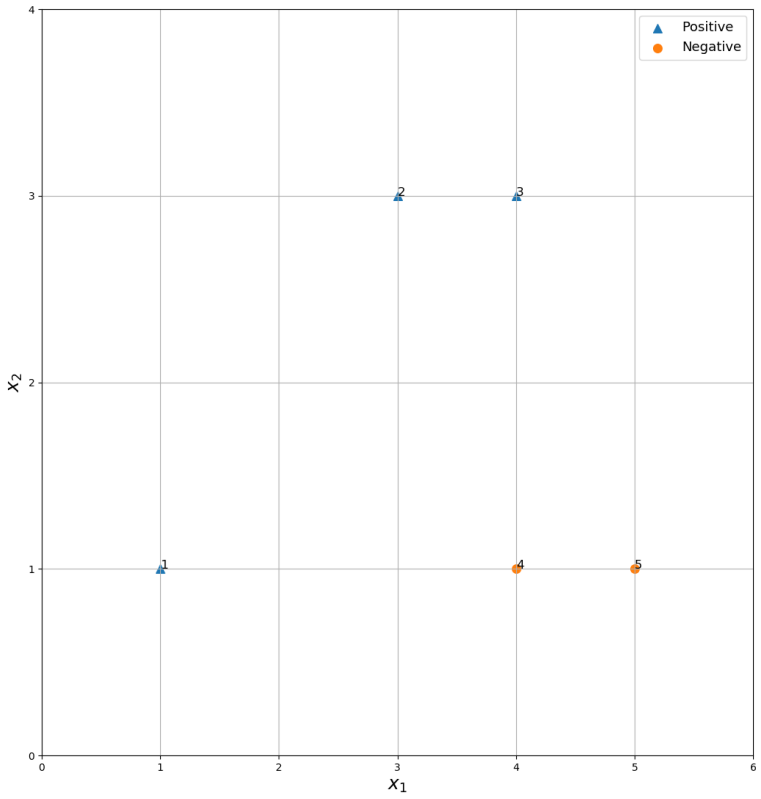


Figure 1: Dataset for AdaBoost

4. (x points) Circle the point that has the highest weight after renormalizing the weight.

5. (x points) What is the weighted classification error of the first classification tree **after** weights are renormalized.

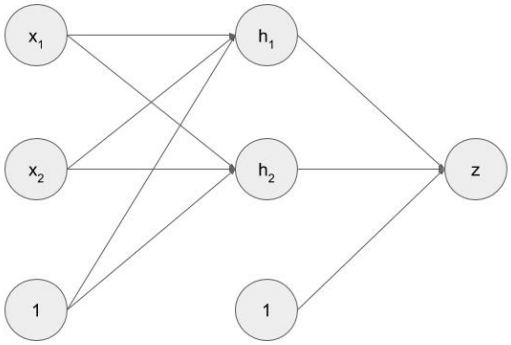
6. (x points) Draw the second classification tree that the algorithm would pick. Again, indicate the positive side and negative side of the decision boundary.

6 [21] Deep Learning

1. Suppose you have a filter of size $k \times k$ and stride s. When you apply this filter to a $n \times n$ input, what is the dimension of the output feature map?

2. Suppose you have
$$input = \begin{bmatrix} 4 & 5 & 1 \\ 4 & 0 & 2 \\ 3 & 4 & 0 \end{bmatrix}$$
and
$$filter = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$
What is the output of applying the filter to the input with stride 1?

3. Consider the following fully connected RELU network.
where
$$\begin{bmatrix} h_1 \\ h_2 \end{bmatrix} = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ 1 \end{bmatrix}$$
$$z = \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix} \begin{bmatrix} f(h_1) \\ f(h_2) \\ 1 \end{bmatrix}$$
$$t = z$$
where $f(h_i) = max(0, h_i)$ and t is the output of the network. Suppose $\begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix} = \begin{bmatrix} 1 & -1 & 1 \\ -1 & -1 & 1 \end{bmatrix}$ and $\begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix} = \begin{bmatrix} 1 & -1 & -2 \end{bmatrix}$. Draw the decision boundary



of the network, namely, $t = 0$ in the range $[-5, 5] \times [-5, 5]$. Please indicate the positive and negative side of the boundary.

4. We are going to use the network in (3) to do regression. Assume that our loss function is

$$l(y, t) = (t - y)^2$$

Show that for a single training example, $x = [x_1, x_2]$

$$\frac{\partial l}{\partial v_i} = 2(t-y)f(h_i) \text{ for } i \neq 3$$
$$\frac{\partial l}{\partial v_3} = 2(t-y)$$
$$\frac{\partial l}{\partial w_{ij}} = 2(t-y)v_i \mathbb{I}(h_i > 0)x_j \text{ for } j \neq 3$$
$$\frac{\partial l}{\partial w_{i3}} = 2(t-y)v_i \mathbb{I}(h_i > 0)$$

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

This page is left blank for scratch space.

This page is left blank for jokes. (Nothing too dirty.)

Please do not write on this page. For administrative purposes only.

T/F	
BV	
Kernels	
CART	
ENSEMBLE	
DL	
TOTAL	