Midterm 3

Printed Name:

Due date:

1. The exam is due Monday 5 Dec 8AM.
2. You may submit it either on sakai electronically or by putting a physical copy under my door.

Rules:

1. The exam is untimed. You do not have to complete the exam in a single sitting. You may pause and restart whenever you'd like.
2. You may use any non-human resources that you like, including notes, books, internet articles, and computers.
3. You are not allowed to discuss the exam in any way with any human until after the due date. This includes:
   (a) obviously bad behavior like copying answers,
   (b) more banal behavior such as:
      i. telling your friend "Problem 6 was really hard" or
      ii. asking your friend "Have you completed the exam yet?"

   Even after you finish the exam, you may not discuss it.
4. If you do have questions about the exam, you should email me the questions rather than posting to github.

Grading:

1. For the True/False/Open questions: Each correct answer will be awarded +1 point, each incorrect answer will result in a -1 point penalty, and each blank answer will result in 0 points.
2. All other problems are worth 1 point, with no penalty for incorrect answers.
3. There are 16 points possible on the exam. Your final grade entered into sakai will be

$$\min\{15, \text{the number of points earned}\}.$$

4. If you find a substantive error on the exam, then I will award you +1 bonus point.

Good luck :)

Problem 1. For each statement below, circle `True` if the statement is known to be true, `False` if the statement is known to be false, and `Open` if the statement is not known to be either true or false. Ensure that you pay careful attention to formal definitions in your responses.

1. `True` `False` `Open` If $|E_{\text{in}} - E_{\text{val}}|$ is large and E_{in} is small, then VC theory predicts that you should decrease the size of your VC dimension.

2. `True` `False` `Open` If you are training a logistic regression model with the polynomial kernel that is overfitting, then VC theory predicts that increasing the degree of the kernel is more likely to improve performance than decreasing the degree of the kernel.

3. `True` `False` `Open` You have trained a logistic regression model with L2 regularization and the polynomial kernel of degree 3. If you increase the degree of the polynomial kernel to 10, then the optimal soft order constraint regularization hyperparameter C will also increase.

4. `True` `False` `Open` When training a neural network with the ReLU activation function and one hidden layer, increasing the width of the hidden layer will increase the generalization error.

5. `True` `False` `Open` You have a dataset with the number of features $d = 10^6$. You have trained a boosted SVM and used a validation set to determine that the optimal number of base classifiers T is 1000. If instead of using an SVM as the base model you use a decision stump, then VC theory predicts that you will need to increase the number of base classifiers T in order to achieve the same generalization error.

6. `True` `False` `Open` The VC dimension of neural networks with the ReLU activation function is $\Omega(d)$, where d is the number of input feature dimensions.

7. `True` `False` `Open` Assume you are training an SVM with the polynomial kernel on a dataset with $N = 10^6$ and $d = 10^6$. You are not using any regularization, and you run the optimization long enough so that optimization error is 0. Then in the limit as the degree of the polynomial approaches infinity, the training error is guaranteed to approach 0 for all possible datasets.

8. `True` `False` `Open` In vowpal wabbit, increasing the `--l1` hyperparameter tends to increase the generalization error.

9. `True` `False` `Open` You have trained a scikit-learn `sklearn.linear_model.LogisticRegression` model with default hyperparameters. It has a high approximation error, low estimation error, and zero optimization error. VC theory predicts that changing the `solver` hyperparameter from the default of `'lbfgs'` to `'saga'` will improve performance.

10. `True` `False` `Open` You are using transfer learning to train the final layer of a deep neural network for your specific task. The `ResNet18` model has 18 hidden layers and the `ResNet50` model has 50 hidden layers, and in both cases all layers have the same width. VC theory predicts that if you use the `ResNet50` model you will have a higher generalization error than if you use the `ResNet18` model.

Problem 2. Provide an example of a hypothesis class implemented by scikit learn with an infinite VC dimension. Also, describe a learning problem where it makes sense to use this hypothesis class.

Problem 3. You work at an online advertisement company that uses vowpal wabbit to determine which ads to display to which users. You have trained a binary classification problem with hyperparameters

```
--bit_precision=28
--ngram=2
--passes=20
--learning_rate=0.01
```

Describe the likely effect on in-sample and generalization errors if you change the value of `--bit_precision` to `22` and `--ngram` to `3` (while keeping all other hyperparameters constant).

Problem 4. You have a dataset with $N = 10^3$ training data points and $d = 10^2$ feature dimensions. You are training a logistic regression model using second order stochastic gradient descent, and you are not computationally limitted so you run the optimization long enough for the optimization error to be negligible. You have evaluated the model on a separate validation set in order to determine that the PCA kernel with output dimension $d' = 10$ and elastic net regularization with $\lambda = 10^{-3}$ and $\alpha = 0.1$ provide good performance. Your boss wants you to simplify the training procedure by removing the PCA kernel feature map. Assuming you keep α constant, how does VC theory predict that you should change λ? That is, do you expect it to increase, decrease, or stay the same? Why?

Problem 5. You have successfully trained a `sklearn.neural_network.MLPClassifier` model and used a validation set to determine that the optimal hyperparameters are `hidden_layer_sizes=[1000]` and `alpha=0.0001`. (All other hyperparameters are the default values.) If you change the value of `hidden_layer_sizes` to `[10000]`, how would you expect to change the `alpha` hyperparameter in order to achieve the same generalization error? That is, do you expect it to increase, decrease, or stay the same? Why?

Problem 6. You are training a boosted decision stump model on a datset with $N = 10^6$ and $d = 10^6$, and have found that the optimal number of decision stumps in the model is $T = 10^3$. If you increase the amount of the data to $N = 10^9$ and the number of features to $d = 10^9$, but keep the number of base models constant at $T = 10^3$, does VC theory predict that the generalization error would increase, decrease, or stay the same?

To justify your answer, you should compute a tight upper bound on the VC dimension of the boosted decision stump hypothesis class and apply the fundamental theorem of statistical learning.

Problem 7. In this problem you will derive a closed form solution to the ridge regression problem, which is closely related to the OLS regression problem. Recall that OLS uses the linear hypothesis class

$$\mathcal{H} = \left\{ \mathbf{x} \mapsto \mathbf{w}^T \mathbf{x} : \mathbf{w} \in \mathbb{R}^d \right\} \tag{1}$$

and the squared loss

$$\ell(\hat{y}, y) = (\hat{y} - y)^2 \tag{2}$$

so that the in-sample error is defined as

$$E_{\text{in}}(h) = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i^T \mathbf{w} - y_i)^2 = \|X\mathbf{w} - \mathbf{y}\|_2^2, \tag{3}$$

where X is a $N \times d$ matrix with ith row equal to $\mathbf{x}_i$ and $\mathbf{y}$ is the d dimensional vector with ith position equal to y_i. Finally, we computed the parameters for the OLS model by solving the equation

$$\hat{\mathbf{w}}^{\text{OLS}} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \|X\mathbf{w} - \mathbf{y}\|_2^2. \tag{4}$$

The ridge regression model modifies Equation (4) above by adding L2 regularization to get

$$\hat{\mathbf{w}}^{\text{ridge}} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \|X\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_2^2. \tag{5}$$

Your task is to derive a closed form solution for $\mathbf{w}^{\text{ridge}}$.