This assignment has 17 questions, for a total of 110 points. Please write your answers **clearly within the space provided** for each question.

I affirm that I have NOT given or received any unauthorized help on this assignment, and that this work is my own.

Your Name: _____

Signature: _____

1. [4 points] Mark each of the following as `T` if the statement is *true* and `F` if the statement is *false*
   (a) _____ If we reduce the size of the training set while keeping the same model complexity, overfitting will be more likely.
   (b) _____ Assume a classifier is trained using gradient descent. As we decrease the number of iterations, overfitting is more likely to happen.
   (c) _____ Duplicating a feature in linear regression does not reduce the mean squared error.
   (d) _____ The squared loss coupled with L2 regularization is a convex function.

2. [4 points] Mark each of the following as `T` if the statement is *true* and `F` if the statement is *false*
   (a) _____ Making a decision tree deeper will likely reduce training error and increase test error.
   (b) _____ If a decision tree performs badly on both training and test sets. It is possible that the tree is too shallow.
   (c) _____ When pruning an already trained decision tree, we usually achieve better test accuracy.
   (d) _____ Using k-fold cross-validation during training will guarantee the model does not overfit.

3. [4 points] Mark each of the following as `T` if the statement is *true* and `F` if the statement is *false*
   (a) _____ LASSO (least absolute shrinkage and selection operator) is a linear regression model where weights are regularized with the ℓ_2-norm.
   (b) _____ The term $\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}$ in the closed form solution for rigde regression is always invertible.
   (c) _____ At every iteration of stochastic gradient descent, we take steps in the exact direction of the gradient vector.
   (d) _____ The solution for $\arg\min_{\mathbf{w}} \quad \frac{1}{n}\|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2+\lambda\|\mathbf{w}\|_2^2$ is not unique.

4. [4 points] Mark each of the following as `T` if the statement is *true* and `F` if the statement is *false*

   (a) _____ LOOCV generally gives more accurate estimates of the test error than 10-fold cross validation.

   (b) _____ A 3-NN classifier is more robust to outliers than 1-NN.

   (c) _____ In logistic regression, if we want to prevent the weights from getting too large we can use a small regularization constant (λ).

   (d) _____ The cost function of logistic regression with L2 regularization is not convex.

5. [4 points] Mark each of the following as `T` if the statement is *true* and `F` if the statement is *false*

   (a) _____ Assuming a logistic regression classifier and a datapoint currently classified as correct, and far away from the decision boundary. If this datapoint is removed, and the classifier retrained, the decision boundary does not change.

   (b) _____ We can expect k-NN to do worse than logistic regression when the data is not linearly separable.

   (c) _____ When the data is not linearly separable, then there is no solution to the hard-margin SVM.

   (d) _____ As the value of C approaches 0, the soft-margin SVM is equal to the hard-margin SVM.

6. Assume we want to minimize the following regularized loss function:

$$L(\mathbf{w}) = \frac{1}{n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{w}\|_1$$

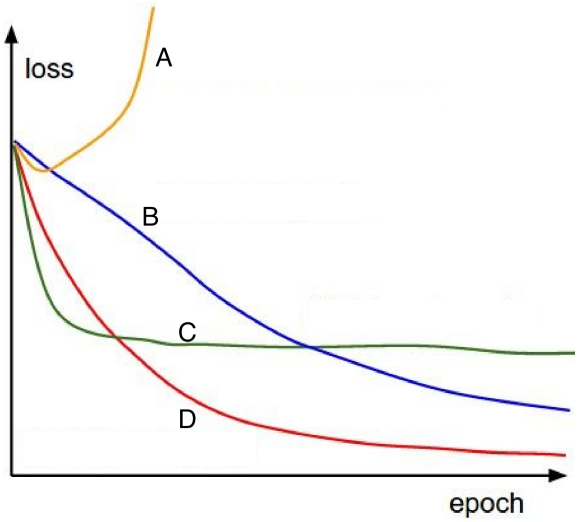
   (a) [5 points] What is the role of the parameter λ?

   (b) [5 points] What happens when $\lambda = 0$?

7. [5 points] For the following dataset, draw a Decision Tree of *minimum depth* that is consistent with the data.

x_1	x_2	y
0	0	-1
0	1	-1
1	0	-1
1	1	+1

8. The figure below shows the loss of training a model using gradient descent.



(a) [5 points] What is the curve that corresponds to the largest learning rate?

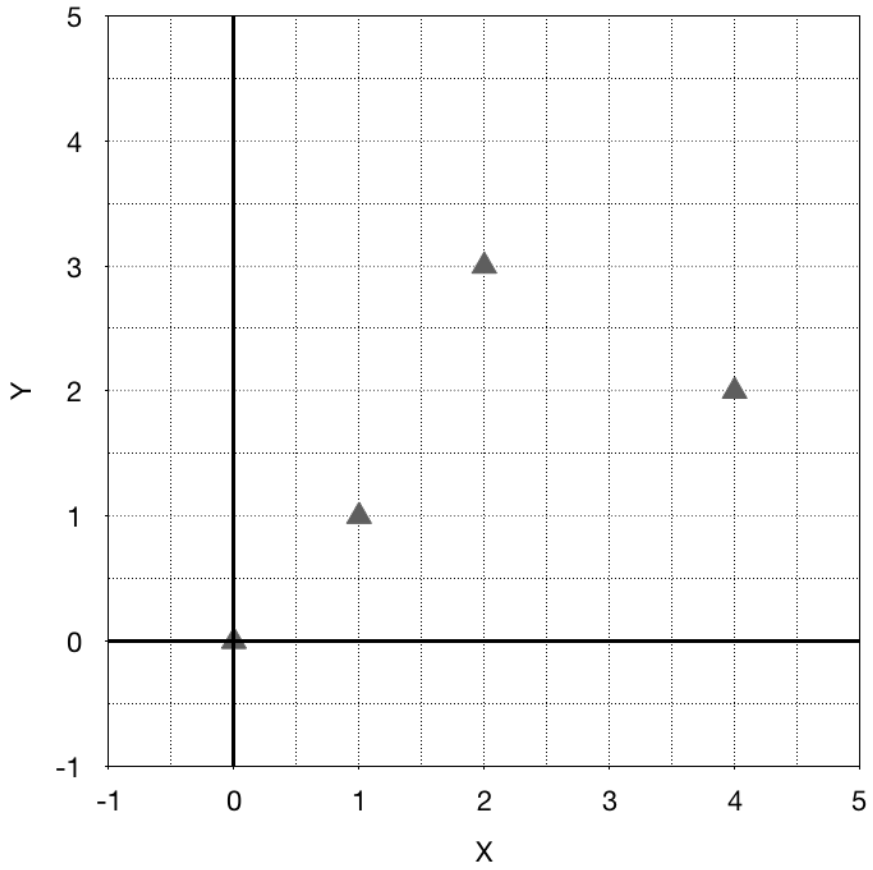
(a) _____

(b) [5 points] What is the curve that corresponds to the smallest learning rate?

(b) _____

9. [5 points] Consider the feature transform $\Phi(x) = [1, \phi_1(x), \phi_2(x)]^T = [1, x, x^2]^T$ and the linear model $h_{\mathbf{w}}(x) = \mathbf{w}^T\Phi(x)$. For the hypothesis with $w = [0.25, -1, 0.5]^T$, what is $h_{\mathbf{w}}(x)$ explicitly as a function of x?

10. Consider the dataset in the chart below. Assume we apply linear regression.



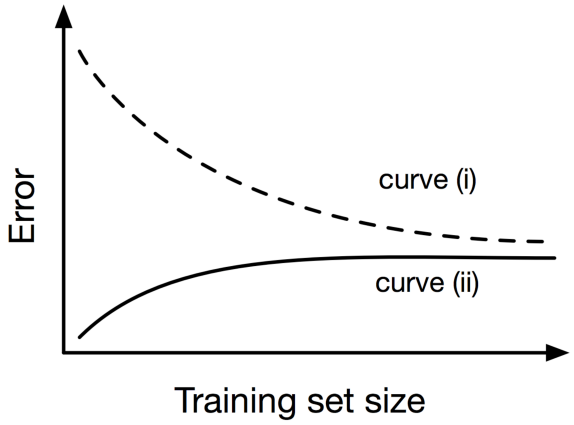
(a) [5 points] What is the MSE when $\mathbf{w}^T = [1, 0]$?

(a) _____

(b) [5 points] What is the MSE when $\mathbf{w}^T = [0.5, 1]$?

(b) _____

11. [5 points] The figure below shows a general trend of how the training and test errors change as we increase the training set size. Which curve best represents the training error? Justify your answer.



12. [5 points] Derive a gradient descent update rule that minimizes the loss function below:

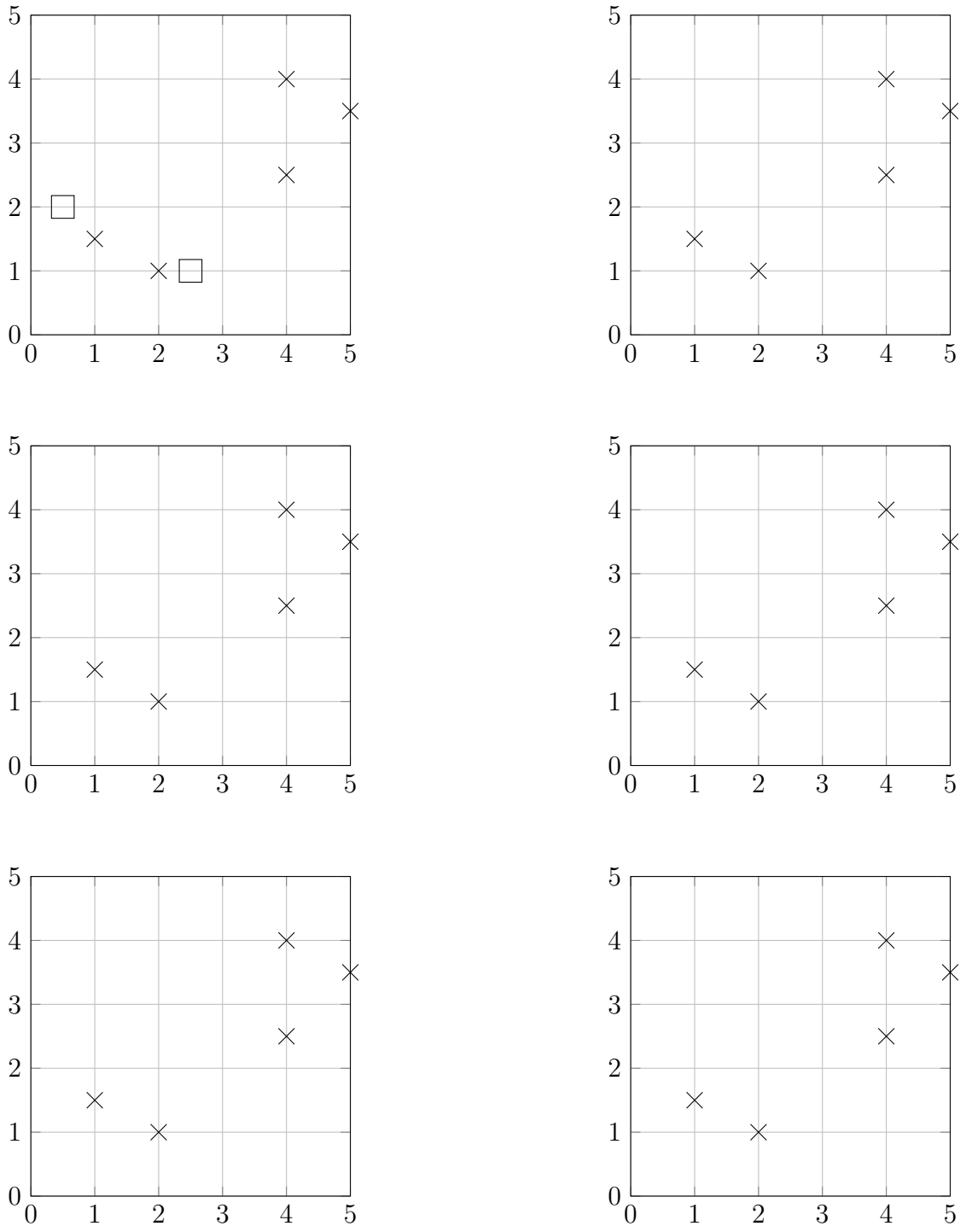
$$L(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\left(h_{\mathbf{w}}(\mathbf{x}^{(i)}) - y^{(i)}\right)^2$$

where:

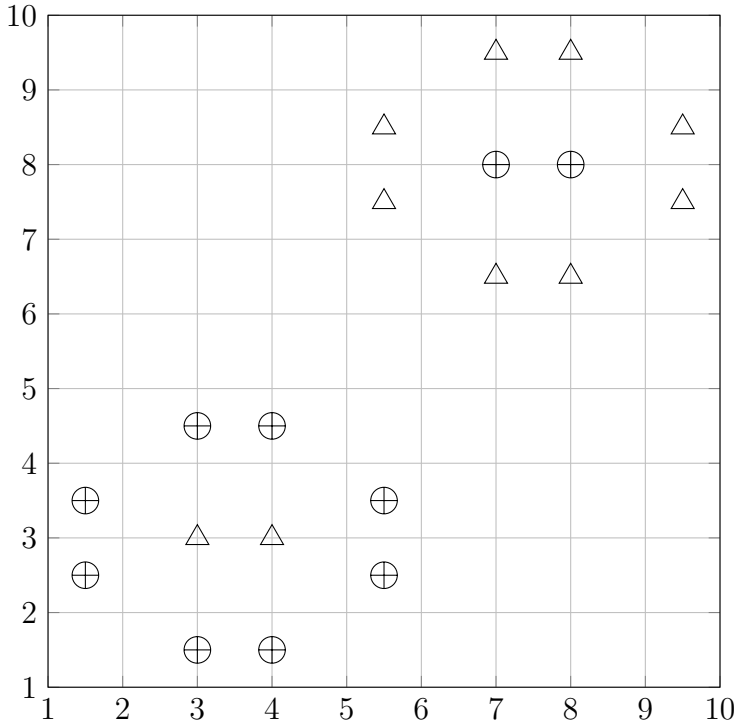
$$h_{\mathbf{w}}(\mathbf{x}) = w_0 + w_1 x_1 + w_1 x_1^3 + w_2 x_2 + w_2 x_2^3 + \cdots + w_d x_d + w_d x_d^3$$

13. [5 points] Consider clustering N 1-dimensional instances on a line situated at coordinates equaling the powers of two $1, 2, 4, 8, \ldots$. Suppose we use hierarchical agglomerative clustering with *single linkage*, i.e. by merging the clusters with the two closest datapoints at each step. Draw the dendrogram obtained after performing clustering on the first 6 points. Clearly indicate the height value of each join.

14. [5 points] Starting with two centroids (squares), perform the k-means algorithm on the data points below (use euclidean distance). In each plot, indicate *data assignments* and starting in the second plot (left-to-right then top-down) draw the new centroids. Stop after convergence or after 5 steps.



15. Considering the dataset below and the use of Euclidean distance:



(a) [5 points] What value of k minimizes the Leave-One-Out Cross-Validation (LOOCV) error for a k-NN classifier? (prefer the smallest possible k)

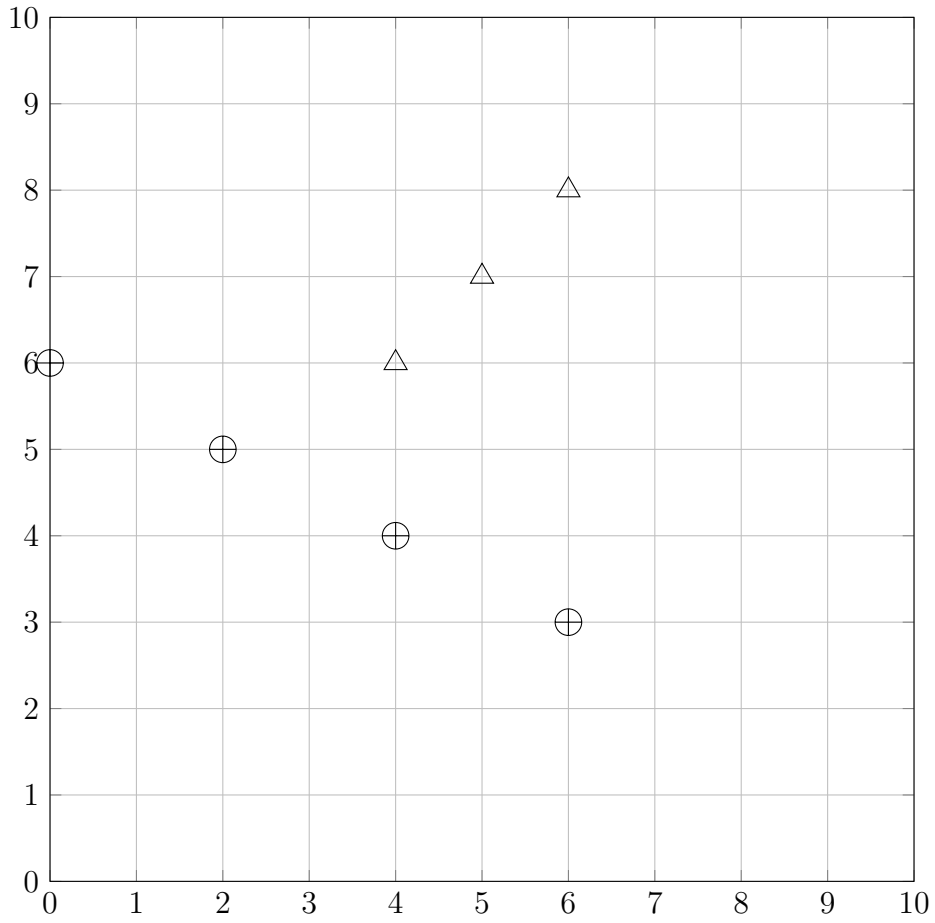
(a) _____

(b) [5 points] What is the final accuracy (between 0% and 100%, inclusive) for such value of k?

(b) _____

16. [5 points] Assuming a linearly separable case, how can we use the solution of the SVM's optimization function to determine which datapoints are the support vectors? (assume $\|\mathbf{w}\| = 1$)

17. Considering applying a linear SVM to the dataset shown below. The final classifier has the form $h_w(\mathbf{x}) = \sigma\left(\mathbf{w}^T\mathbf{x} + b\right)$, with $\|w\| = 1$



(a) [5 points] What are the final values of $\mathbf{w}$ and b?

(b) [5 points] Indicate the equation corresponding to the decision boundary

---

(c) [5 points] What is the 7-fold cross validation accuracy (between 0% and 100%, inclusive)? Note that the decision boundary is recalculated for each fold.
