Linear Regression (Dani)

1. Optimal MSE rule [10 pts]

Suppose we knew the joint distribution P_{XY}. The optimal rule $f^* : X \to Y$ which minimizes the MSE (Mean Square Error) is given as:

$$f^* = \arg\min_f \mathbb{E}[(f(X) - Y)^2]$$

Show that $f^*(X) = \mathbb{E}[Y|X]$.

(10 points)
Notice that it suffices to argue that

$$\mathbb{E}[(f(X) - Y)^2] \geq \mathbb{E}[(\mathbb{E}[Y|X] - Y)^2] \text{ for all } f$$

and hence $f^*(X) = \mathbb{E}[Y|X]$.

$$\begin{aligned}
\mathbb{E}[(f(X) - Y)^2] &= \mathbb{E}[(f(X) - \mathbb{E}[Y|X] + \mathbb{E}[Y|X] - Y)^2] \\
&= \mathbb{E}[(f(X) - \mathbb{E}[Y|X])^2 + (\mathbb{E}[Y|X] - Y)^2 + 2(f(X) - \mathbb{E}[Y|X])(\mathbb{E}(Y|X) - Y)] \\
&= \mathbb{E}[(f(X) - \mathbb{E}[Y|X])^2] + \mathbb{E}[(\mathbb{E}[Y|X] - Y)^2] + 2\mathbb{E}[(f(X) - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - Y)]
\end{aligned}$$

Now using the fact that $\mathbb{E}_{XY}[\dots] = \mathbb{E}_X[\mathbb{E}_{Y|X}[\dots|X]]$, we have

$$\begin{aligned}
\mathbb{E}_{XY}[(f(X) - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - Y)] &= \mathbb{E}_X[\mathbb{E}_{Y|X}[(f(X) - \mathbb{E}[Y|X])(\mathbb{E}[Y|X] - Y)|X]] \\
&= \mathbb{E}_X[(f(X) - \mathbb{E}[Y|X])\mathbb{E}_{Y|X}[(\mathbb{E}[Y|X] - Y)|X]] = 0
\end{aligned}$$

where the second last step follows since conditioning on X, $f(X)$ and $\mathbb{E}[Y|X]$ are constant. Therefore,

$$\begin{aligned}
\mathbb{E}[(f(X) - Y)^2] &= \mathbb{E}[(f(X) - \mathbb{E}[Y|X])^2] + \mathbb{E}[(\mathbb{E}[Y|X] - Y)^2] \\
&\geq \mathbb{E}[(\mathbb{E}[Y|X] - Y)^2]
\end{aligned}$$

since the first term being square of a quantity is non-negative.
Note: There are many ways to prove this. Please use your judgement and give partial credit if some arguments are right and others are not.

5 Regression with Regularization [10 Points]

You are asked to use regularized linear regression to predict the target $Y \in \mathbb{R}$ from the eight-dimensional feature vector $X \in \mathbb{R}^8$. You define the model $Y = w^T X$ and then you recall from class the following three objective functions:

$$\min_w \sum_{i=1}^{n} \left(y_i - w^T x_i\right)^2 \tag{5.1}$$

$$\min_w \sum_{i=1}^{n} \left(y_i - w^T x_i\right)^2 + \lambda \sum_{j=1}^{8} w_j^2 \tag{5.2}$$

$$\min_w \sum_{i=1}^{n} \left(y_i - w^T x_i\right)^2 + \lambda \sum_{j=1}^{8} |w_j| \tag{5.3}$$

1. **[Points: 2 pts]** Circle regularization terms in the objective functions above.

   ★ **SOLUTION:** The regularization term in 5.2 is $\lambda \sum_{j=1}^{8} w_j^2$ and in 5.3 is $\lambda \sum_{j=1}^{8} |w_j|$.

2. **[Points: 2 pts]** For large values of λ in objective 5.2 the bias would:

   (a) increase ★
   (b) decrease
   (c) remain unaffected

3. **[Points: 2 pts]** For large values of λ in objective 5.3 the variance would:

   (a) increase
   (b) decrease ★
   (c) remain unaffected

4. **[Points: 4 pts]** The following table contains the weights learned for all three objective functions (not in any particular order):

	Column A	Column B	Column C
w_1	0.60	0.38	0.50
w_2	0.30	0.23	0.20
w_3	-0.10	-0.02	0.00
w_4	0.20	0.15	0.09
w_5	0.30	0.21	0.00
w_6	0.20	0.03	0.00
w_7	0.02	0.04	0.00
w_8	0.26	0.12	0.05

   Beside each objective write the appropriate column label (A, B, or C):

   - Objective 5.1: ★ **Solution:** A
   - Objective 5.2: ★ **Solution:** B
   - Objective 5.3: ★ **Solution:** C

6 Controlling Overfitting [6 Points]

We studied a number of methods to control overfitting for various classifiers. Below, we list several classifiers and actions that might affect their bias and variance. Indicate (by circling) how the bias and variance change in response to the action:

1. [**Points: 2 pts**] Reduce the number of leaves in a decision tree:

   ★ **SOLUTION:**

Bias	Variance
Decrease	Decrease ★
★ Increase	Increase
No Change	No Change

2. [**Points: 2 pts**] Increase k in a k-nearest neighbor classifier:

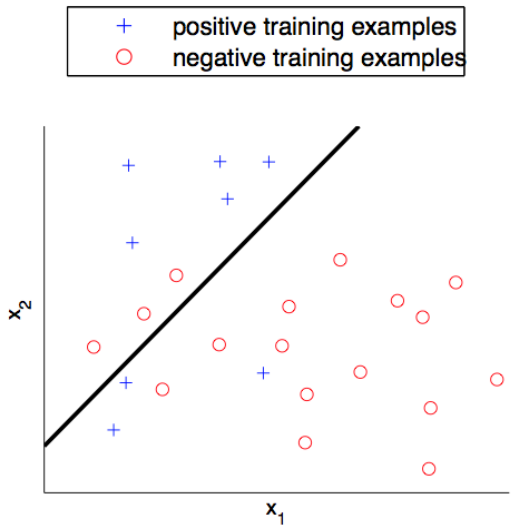
Bias	Variance
Decrease	Decrease ★
★ Increase	Increase
No Change	No Change

3. [**Points: 2 pts**] Increase the number of training examples in logistic regression:

Bias	Variance
Decrease	Decrease ★
Increase	Increase
★ No Change	No Change

7 Decision Boundaries [12 Points]

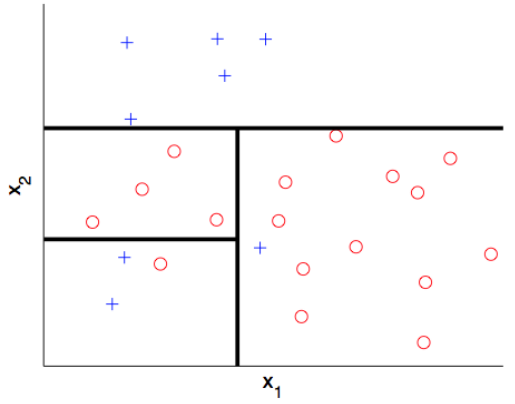
The following figures depict decision boundaries of classifiers obtained from three learning algorithms: decision trees, logistic regression, and nearest neighbor classification (in some order). Beside each of the three plots, write the **name** of the learning algorithm and the **number of mistakes** it makes on the training data.



[Points: 4 pts]

Name: ★ Logistic regression

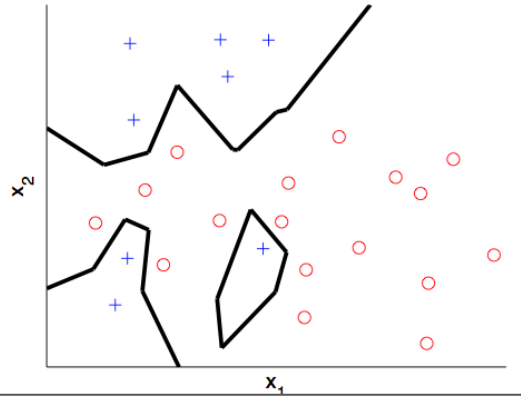
Number of mistakes: ★ 6



[Points: 4 pts]

Name: ★ Decision tree

Number of mistakes: ★ 2



[Points: 4 pts]

Name: ★ k-nearest neighbor

Number of mistakes: ★ 0