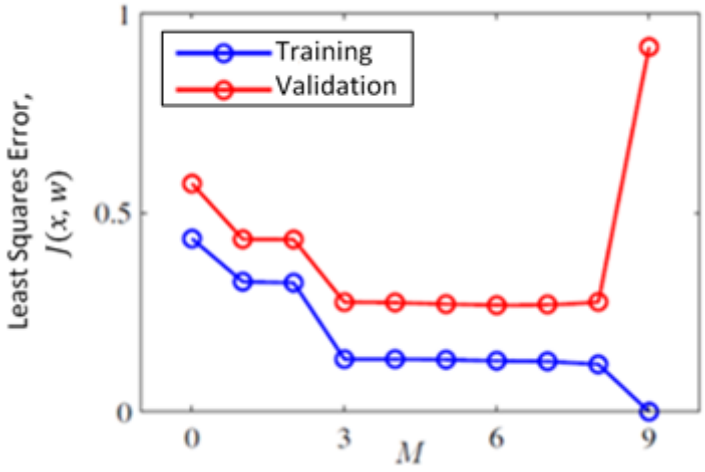
Practice Problems

Problem 1 – Polynomial Regression

Consider the polynomial curve fitting example discussed in class:

$$y(x, \boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=1}^{M} w_j x^j$$

Suppose you have a data set, $\boldsymbol{X}$, that you split into two groups once: a fixed training subset, X_train, and a fixed validation subset $X\_\text{validation}$. After fitting the polynomial to the training set across a range of model orders and evaluating on both training and validation sets, you obtain the following plot.



Answers the following questions:

1. Based on this plot, provide a discussion about which model order, M, should be used to avoid overfitting.
2. Since you have split your data into training and validation sets, are you safe from overfitting or can you still overfit? Explain your reasoning.
3. What are the common approaches to avoid overfitting given a model?
4. For the data set $\boldsymbol{X} = \{x_1, x_2, \ldots, x_N\}$ and the labels $\boldsymbol{t} = \{t_1, t_2, \ldots, t_N\}$, what objective function can you use if you want to optimize the polynomial linear regression model with a regularizer on the weights $\boldsymbol{w} = [w_1, w_2, \ldots, w_M]^T$? Write down the equation for this objective function.

Problem 2 – Curse of Dimensionality

What is the curse of dimensionality? Why does it cause problems for certain Machine Learning algorithms? Describe two approaches to address curse of dimensionality. In your description state why each method is effective and what are its limitations.

Problem 3 – Curse of Dimensionality

Recall our discussion of the volume of the crust, i.e., the case of a sphere S_2 of radius $r - \epsilon$ inscribed within another sphere S_1 of radius r and the relative volume of the crust and the outer sphere as we increase dimensionality D:

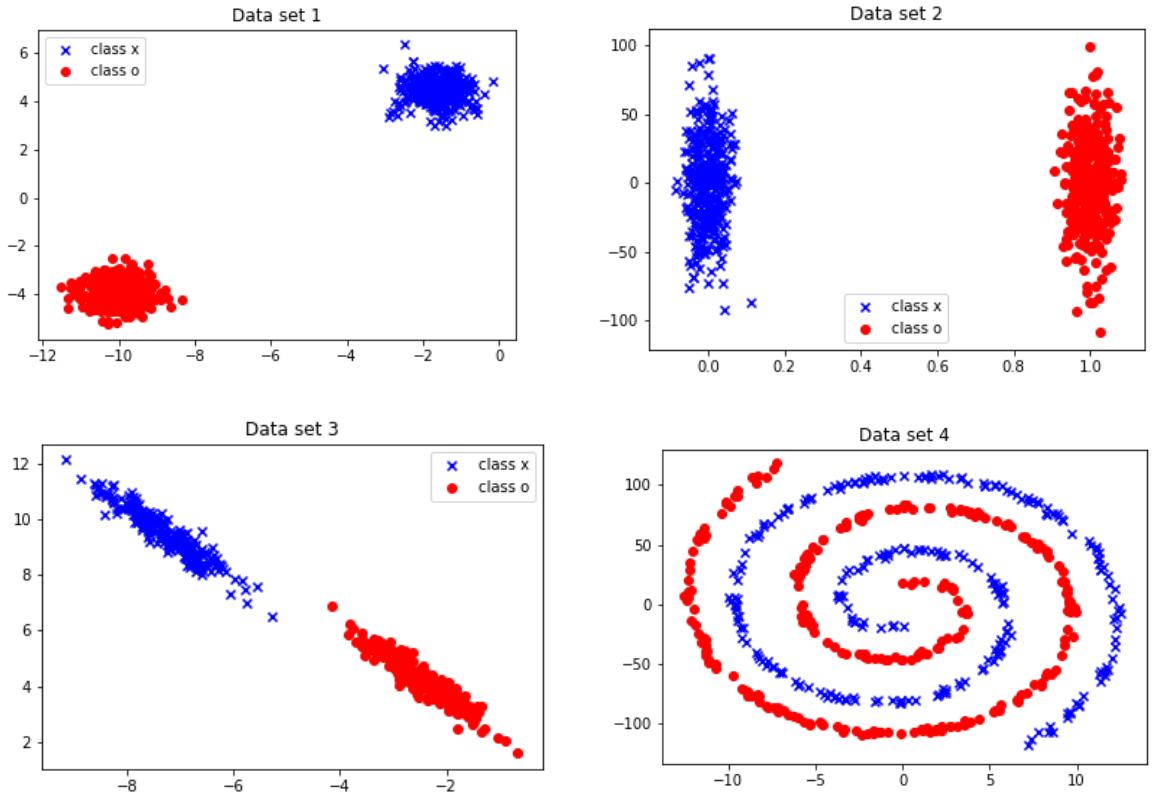
$$\frac{V_{\text{crust}}}{V_{S_1}} = \frac{V_{S_1} - V_{S_2}}{V_{S_1}} = 1 - \left(1 - \frac{\epsilon}{r}\right)^D$$

Describe (in paragraph form, be clear and thorough) how this concept relates to number of data points needed during classification as we increase the number of features we use for classification. In your discussion, be sure to answer: (1) the case when feature are uncorrelated, (2) the case when features are strongly correlated (e.g., $f_1 = af_2 + bf_3$), and (3) why this is an important issue in machine learning in general.

Problem 4 – Principal Component Analysis

Consider the following four two-dimensional data sets each containing two clusters of data points (shown with "circles" and "crosses"). Suppose you would like to apply Principal Component Analysis (PCA) to reduce the dimensionality of each of these data sets from 2-D to 1-D where the two clusters remain separated in the 1-D projection. For each data set, address each of the following questions:

1. Draw the eigenvectors on each figure and clearly identify the (axis) direction of 1-D PCA projection.
2. Will PCA be effective at keeping the two clusters separated in the 1-D projection? Why or why not? If yes, state what characteristics of the data set allow PCA to be effective. If no, state what characteristics of the data set cause PCA to fail.
3. Can you think of another technique that would be successful at reducing the dimensionality of this data set while maintaining (or increase) separation between the two clusters? State the other method and describe why it would be successful.



Problem 5 – Principal Component Analysis

Suppose $\boldsymbol{X}$ is a zero-mean data set of size $2 \times N$ with covariance matrix $R_X = \begin{bmatrix} 1 & 0 \\ 1 & 3 \end{bmatrix}$.

1. Write down the eigenvalue/eigenvector equation.
2. Calculate the eigenvalues λ_1 and λ_2 associated with the eigenvectors $u_1 = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$ and $u_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, respectively, of R_X.
3. Suppose $\boldsymbol{Y}$ is the PCA transformation of $\boldsymbol{X}$. Write the formula for computing $\boldsymbol{Y}$ from $\boldsymbol{X}$. Be as specific as possible, you may use information from earlier parts.
4. What is the covariance matrix of $\boldsymbol{Y}$? Why? Be as specific as possible, you may use information from earlier parts.
5. What is the amount of explained variance of the 1-D PCA projection?

Problem 6 – Cross-Validation

Suppose you have 100 training samples that you are using to train a classifier to distinguish between four classes. The training data has 50 samples of class 1, 25 samples of class 2, 20 samples of class 3 and 5 samples of class 4. To evaluate the stability and performance of your classifier on each class, you use 10-fold cross-validation. Is it a good strategy to randomly partition the data into 10 folds? Why or why not? If yes, fully justify why. If no, state why not, provide an alternate cross-validation scheme and justify the new scheme.

Problem 7 – Regularization

Suppose you would like to perform feature selection by beginning with too many extracted features and, then, driving the weights for unnecessary features to zero using an appropriate regularization term. Considering the following two regularization terms,

$$E_{c,1} = \sum_{i=1}^{M} w_i^2$$

$$E_{c,2} = \sum_{i=1}^{M} |w_i|$$

which would be more effective for driving the weights to zero? Why? Clearly explain your reasoning.

Problem 8 – ROC Curves and Confusion Matrices

Suppose you have the following training data set (associated with data collected during 1 minute of a detection system):

$$X = \{(1,1,2);\ (10,3,0);\ (-5,-4,1);\ (2,-3,1);\ (10,10,20);\ (0,0,0)\}$$

$$y = \{-1, 1, 1, -1, -1, 1\}$$

where $y_i = 1$ indicates a true target and $y_i = -1$ indicates a non-target data point. Suppose you trained a classifier to produce a confidence of target given a sample. For the above data points, your MLP produced the following confidence values:

$$c = \{0.7,0.6,0.2,0.3,0, 0.9\}$$

1. Draw the associated ROC curve.
2. To make a final decision, suppose you use a threshold at a confidence value of 0.5 where everything ≤ 0.5 is marked as non-target and > 0.5 is marked as target. What would be your resulting confusion matrix?

Problem 9 – Learning System

Draw the block diagram for a learning system and explain in words the function of each block.

Problem 10 – PCA vs Regression

What is the difference between PCA and regression? Be specific by explaining what each method accomplishes in terms of data representation. Illustrate with an example of a problem where you will use PCA and also another where you will use regression.

Problem 11 – Dimensionality Reduction

Suppose you would like to use PCA to reduce dimensionality of your data prior to applying a k-NN classifier. Is PCA always an effective dimensionality reduction technique to be used in conjunction with classification? Why or why not?

Problem 12 – Decision Tree vs Random Forest

Describe (using pseudo-code and precise descriptions) the difference between a decision tree and a random forest. Focus your discussion on how the decision trees and random forests are constructed/trained.

Problem 13 – PCA vs LDA

Describe the differences between PCA and LDA. Illustrate each algorithm with an example.