Midterm Questions

1. [4 points] What is the difference in purpose between a training data set and a validation data set?

   Training data set is used for learning the parameters of the models. Validation data set is used to decide which model to employ. The latter is used to find the best hyper parameters (like learning rate, regularization strength etc.)

   [2+ 2, with partial points]

2. [4 points] (a) What is a regression problem? (b) Is it an example of supervised or unsupervised learning? (c) Does it provide a descriptive or a predictive model?

   (a) It is the statistical process for estimating the relationships among variables and it involves learning a real-valued function from training labeled data.

   (b) It is an example of supervised learning since it requires training from labeled data.

   (c) It is a predictive model.

   [2+1+1, no partial points for b and c]

3. [6 points] Points A=(6, 0) and B=(0, 6) are equidistant from point C=(0, 0) in feature space, using a Euclidian distance measure. What point D=(*d, d*) is also equidistant from C using (a) L1 distance measure, (b) L2 distance measure, and (c) L_∞ distance measure? (Give the value of *d* for each.)

   a. Under the L1 distance measure, D=(3,3) is the same distance from C
   b. Under the L2 distance measure D($\sqrt{18}, \sqrt{18}$) is equidistant from C
   c. Under the L_∞ measure, D(6,6) is equidistant from C

   [2 points each, 1 point for explanation/ showing the calculation, 1 for final answer]

4. [4 points] You create a basic linear binary classifier that is consistent but not complete. Out of 50 training samples, what can you determine (if anything) about the number of false positives and the number of false negatives?

   As classifier is consistent => FP =0, There are zero false positives.
   It is not complete => TP +FN = P where FN is non zero. Therefore there is at least one FN.

   [2 + 2, no partial points if you get FP=0 wrong]

5. [8 points] Running a binary classification model on testing data set results in 40 out of the 100 testing instances being misclassified. The true positive rate of the test is

0.50, with 60 positive examples. (a) What is the false positive rate? (b) The false negative rate? (c) The precision? (d) The accuracy?

P=60
N=Total - P = 40
TPR = TP/P = TP/60 =0.5
=> TP = 30
TN = 30

FP = N-TN =40-30=10
FN = P-TP = 60-30=30

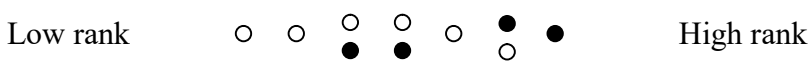
(a) FPR=FP/N=10/40=0.25
(b) FNR=FN/P=30/60=0.5
(c) Precision = TP/~P = TP/ (TP+FP) = 30/(30+10)=0.75
(d) Accuracy = (TP + TN) / Total = (30+30)/100=0.6
[2 points each, 1 for calculation and 1 for final answer]

6. [4 points] Someone proposes this loss function for a regression problem: $L(x) = {}^{1}/_{(R(x))^2}$, where $R(x)$ is the residual. Explain why this is a good or bad loss function for regression.

With a loss function of ${}^{1}/_{(R(x))^2}$ the regression function will perform very poorly because it will apply an infinite penalty to nearly correct results, while applying minimal error to large errors. Additionally it treats positive and negative residuals the same which also creates some issues.

[4 points, partial points based on the answer]

7. [6 points] A ranking classifier produces the following ranking, where filled-in circles are in the positive class and open circles are in the negative class:



There are three pairs that are tied in their rankings. What are the error rate and accuracy of this ranking output?

P = 4, N =6, #Ranking Errors =2.5, 2, 1.5, 0.5 = 6.5

Error Rate = 6.5/(4 x 6) = 0.27

Accuracy = 1-0.27 = 0.73
[2 points each, 1 for calculation and 1 for final answer]

8. [5 points] We have a binary classification problem where there are three features (season of the year, day of the week, and general time of day) that have 4, 7, and 5 possible values, respectively. (a) How many possible hypotheses are there for this problem? (b) If we give numeric values to each feature, how many possible linear classifiers are there?

   (a) There are 4*7*5 possible combinations, therefore, the hypothesis space for binary classification is $2^{(4*7*5)} = 2^{140}$

   (b) As features take numeric values, the range of each of these features can vary up to infinity which implies you can find infinitely many planes in a 3D space. Therefore there are infinitely many possible linear classifiers for this classification task.
   [2+3, partial points for b based on answer]

9. [5 points] In the binary classification case described in the previous problem, if you use a conjunctive hypothesis space (CHS) representation (no internal disjunctions), what is the most general hypothesis that includes the hypothesis (Season=Fall $\wedge$ Day=Sunday)?

   H=True is the most general hypothesis satisfying the condition
   It can also be written as H (X, X, X)
   [5 points with partial points]

10. [5 points] In the binary classification case described in the previous two problems, you are given three conjunctive hypothesis space (CHS) training samples:

    Positive: (Season=Fall $\wedge$ Day=Sunday $\wedge$ Time=EarlyMorning)
    Positive: (Season=Fall $\wedge$ Day=Wednesday $\wedge$ Time=EarlyMorning)
    Negative: (Season=Fall $\wedge$ Day=Sunday $\wedge$ Time=Midday)

    What is the most general CHS hypothesis (no internal disjunctions) that is consistent with the training examples?

    C1 = Season =Fall $\wedge$ Day = [Sunday, Wednesday] $\wedge$ Time = Early Morning
    This covers both the positive examples and no negative example
    However, this is not the most general CHS. [*This will not fetch full points, 3 points*]

    We can refine it further to C2 = Time = EarlyMorning in order to make it the Most General CHS that is consistent with the training samples.
    [5 points, with partial points]

11. [6 points] Using the GrowTree algorithm, build a decision tree from the following training samples that use three features, each of which has two values:

    Positive: (F1=true, F2=false, F3=true)
    Positive: (F1=false, F2=false, F3=true)
    Positive: (F1=false, F2=false, F3=false)
    Negative: (F1=true, F2=true, F3=true)
    Negative: (F1=false, F2=true, F3=false)
    Negative: (F1=true, F2=true, F3=false)

    Split data on F2. Done.
    [6 points, with partial points]

12. [5 points] In a linear least-squares regression problem, we have the following training data in the form of { feature values, label }:

    { (4, 3, 1), 7 }
    { (2, 0, 2), 5 }
    { (-5, -1, 6), 0 }
    { (2, -1, -3), -2 }
    { (0, 0, 1), 3 }

    (a) Write out (provide the numbers, but don't solve) the matrix equation $\widehat{\boldsymbol{w}} = (\boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}})^{-\mathbf{1}}\boldsymbol{X}\boldsymbol{y}$ that finds the regressor $\widehat{\boldsymbol{w}}$.

    (b) For a new input value (3, 3, 3), how do you compute the regression estimate $\widehat{\boldsymbol{y}}$? (Don't solve for $\widehat{\boldsymbol{y}}$, just show how.)

a)

$$\hat{w} = \left(\begin{bmatrix} 4 & 2 & -5 & 2 & 0 \\ 3 & 0 & -1 & -1 & 0 \\ 1 & 2 & 6 & -3 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 4 & 3 & 1 & 1 \\ 2 & 0 & 2 & 1 \\ -5 & -1 & 6 & 1 \\ 2 & -1 & -3 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 4 & 2 & -5 & 2 & 0 \\ 3 & 0 & -1 & -1 & 0 \\ 1 & 2 & 6 & -3 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 7 \\ 5 \\ 0 \\ -2 \\ 3 \end{bmatrix}$$

b) Write the equation for $\widehat{\boldsymbol{y}}$,

$$\hat{y} = \hat{w}^T \begin{bmatrix} 3 \\ 3 \\ 3 \\ 1 \end{bmatrix}$$

Note that a solution which is not extended with 1's can only produce lines centered around the origin.
[3+2, with partial points]

13. [4 points] Using the Gini index, what is the total impurity of a node (in a feature tree) with three attribute values that partitions the training data in this way:

    Node feature = value 1 : 10 positives, 5 negatives
    Node feature = value 2 : 2 positives, 6 negatives
    Node feature = value 3 : 4 positives, 4 negatives

    Gini Index Imp ($\dot{p}$) = $2\dot{p}(1-\dot{p})$

    $\dot{p}$ = P/P+N = 10/15 = 0.67 for value 1, Gini Index = 2x0.67x0.33 = 0.4422

    = 2/8 = 0.25 for value 2, Gini Index = 2x0.25x0.75 = 0.375

    = 4/8 = 0.50 for value 3, Gini Index = 2x0.50x0.50 = 0.5

    Gini Index of Node is weighted sum of the above,
    (0.4422x15 + 8x0.375 +8 x0.5)/31
    (6.633 + 3 + 4)/31
    0.4397
    [3+1, with partial points for correct application of formula]

14. [4 points] The linear model at the core of multivariate linear regression is $\boldsymbol{y = Xw + \epsilon}$. For data set of 100 3D data points, give the dimensionality (rows × cols) of the variables $\boldsymbol{y, X, w,}$ and $\boldsymbol{\epsilon}$.

    y: (100x1),
    X: (100x3)
    w: (3x1)
    $\boldsymbol{\epsilon}$: (100x1)

    [1+1+1+1]

15. [4 points] Our basic binary linear classifier (with the classification boundary halfway between the class centroids) is simple and straightforward to create. Learning a basic perceptron classifier is a bit more complicated – it requires iterating on the training data, updating the classification boundary ($\boldsymbol{w}$) for each misclassified training sample. Which of these classifiers (basic linear or basic perceptron) should you use when you know your data is noisy and there is overlap between the two classes? Explain your answer.

    Perceptron Learning will achieve perfect separation on linearly separable data. But as we know that the data is noisy and there is overlap between classes the perceptron learning will not converge. Therefore we should use a linear binary classifier.

    [2 points for mentioning the properties of the two classifiers + 2 for final answer]