Assignment 2 Solution

Instructor: Oliver Schulte

Theory: Variance and Covariance (10 Marks)

For definitions and notation please refer to the text. We write $var(X)$ for the variance of a single random variable and $cov(X, Y)$ for the covariance of two random variables, such that $var(X) = cov(X, X)$.

1. Show that $var(X) = E(X^2) - [E(X)]^2$.

$$\begin{aligned} var(X) &= E((X-\mu)^2) \\ &= E((X-E(X))^2) \\ &= E(X^2 - 2XE(X) + [E(X)]^2) \\ &= E(X^2) - 2E(X)E(X) + [E(X)]^2 \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$

2. Show that if two random variables X and Y are independent, then their covariance is zero.

$$\begin{aligned} cov(X,Y) &= E((X-E(X))(Y-E(Y)) \\ &= E(XY - XE(Y) - YE(X) + E(X)E(Y)) \\ &= E(XY) - E(X)E(Y) - E(Y)E(X) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

X and Y are independent random variables, So,

$$\begin{aligned} cov(X,Y) &= E(XY) - E(X)E(Y) \\ &= E(X)E(Y) - E(X)E(Y) \\ &= 0 \end{aligned}$$

Practice: Decision Tree Learning (15 Marks)

1. Install a package that implements the ID3 decision tree algorithm that we studied in class, for both discrete and continuous input features. We recommend using Weka, see course web page.

```
CSS_rank <= 12
|   rs_G <= 15
|   |   country_group = EURO
|   |   |   rs_PlusMinus <= 0
```

```
|   |   |   |   rs_PlusMinus <= -2: no (3.0)
|   |   |   |   rs_PlusMinus > -2: yes (15.0/1.0)
|   |   |   rs_PlusMinus > 0: no (8.0)
|   |   country_group = CAN: yes (5.0)
|   |   country_group = USA: yes (5.0/1.0)
|   rs_G > 15: yes (36.0)
CSS_rank > 12
|   rs_P <= 11: no (99.0/16.0)
|   rs_P > 11
|   |   rs_PlusMinus <= 0
|   |   |   rs_PlusMinus <= -1
|   |   |   |   country_group = EURO: no (32.0/2.0)
|   |   |   |   country_group = CAN
|   |   |   |   |   rs_PlusMinus <= -18: no (6.0)
|   |   |   |   |   rs_PlusMinus > -18
|   |   |   |   |   |   Position = D
|   |   |   |   |   |   |   po_G <= 2
|   |   |   |   |   |   |   |   Weight <= 198: no (2.0)
|   |   |   |   |   |   |   |   Weight > 198: yes (11.0/2.0)
|   |   |   |   |   |   |   po_G > 2: no (2.0)
|   |   |   |   |   |   Position = C
|   |   |   |   |   |   |   rs_PIM <= 45: yes (3.0)
|   |   |   |   |   |   |   rs_PIM > 45: no (6.0)
|   |   |   |   |   |   Position = L
|   |   |   |   |   |   |   po_GP <= 14
|   |   |   |   |   |   |   |   rs_G <= 12
|   |   |   |   |   |   |   |   |   rs_A <= 10: yes (2.0)
|   |   |   |   |   |   |   |   |   rs_A > 10: no (3.0)
|   |   |   |   |   |   |   |   rs_G > 12: yes (8.0)
|   |   |   |   |   |   |   po_GP > 14: no (3.0)
|   |   |   |   |   |   Position = R
|   |   |   |   |   |   |   Weight <= 215: no (6.0/1.0)
|   |   |   |   |   |   |   Weight > 215: yes (4.0)
|   |   |   |   country_group = USA: yes (11.0/3.0)
|   |   |   rs_PlusMinus > -1
|   |   |   |   CSS_rank <= 39: yes (34.0/3.0)
|   |   |   |   CSS_rank > 39
|   |   |   |   |   rs_GP <= 32
|   |   |   |   |   |   rs_G <= 4: yes (3.0)
|   |   |   |   |   |   rs_G > 4
```

```
|   |   |   |   |   |   |   Height <= 71
|   |   |   |   |   |   |   |   Weight <= 191: yes (2.0)
|   |   |   |   |   |   |   |   Weight > 191: no (2.0)
|   |   |   |   |   |   |   Height > 71: no (15.0)
|   |   |   |   |   rs_GP > 32
|   |   |   |   |   |   country_group = EURO
|   |   |   |   |   |   |   Weight <= 197: yes (12.0)
|   |   |   |   |   |   |   Weight > 197
|   |   |   |   |   |   |   |   CSS_rank <= 86: yes (5.0)
|   |   |   |   |   |   |   |   CSS_rank > 86
|   |   |   |   |   |   |   |   |   CSS_rank <= 234
|   |   |   |   |   |   |   |   |   |   rs_PIM <= 50: yes (3.0/1.0)
|   |   |   |   |   |   |   |   |   |   rs_PIM > 50: no (6.0)
|   |   |   |   |   |   |   |   |   CSS_rank > 234: yes (4.0)
|   |   |   |   |   |   country_group = CAN
|   |   |   |   |   |   |   Height <= 74
|   |   |   |   |   |   |   |   DraftAge <= 18
|   |   |   |   |   |   |   |   |   po_GP <= 13
|   |   |   |   |   |   |   |   |   |   rs_G <= 5: yes (2.0)
|   |   |   |   |   |   |   |   |   |   rs_G > 5: no (9.0/1.0)
|   |   |   |   |   |   |   |   |   po_GP > 13: yes (2.0)
|   |   |   |   |   |   |   |   DraftAge > 18
|   |   |   |   |   |   |   |   |   CSS_rank <= 108: yes (8.0)
|   |   |   |   |   |   |   |   |   CSS_rank > 108
|   |   |   |   |   |   |   |   |   |   Height <= 73
|   |   |   |   |   |   |   |   |   |   |   rs_PIM <= 109: yes (13.0/3.0)
|   |   |   |   |   |   |   |   |   |   |   rs_PIM > 109: no (4.0)
|   |   |   |   |   |   |   |   |   |   Height > 73: no (3.0)
|   |   |   |   |   |   |   Height > 74: no (3.0)
|   |   |   |   |   |   country_group = USA
|   |   |   |   |   |   |   rs_A <= 45
|   |   |   |   |   |   |   |   Height <= 72: yes (18.0/3.0)
|   |   |   |   |   |   |   |   Height > 72
|   |   |   |   |   |   |   |   |   rs_PIM <= 62: yes (11.0/4.0)
|   |   |   |   |   |   |   |   |   rs_PIM > 62: no (11.0/3.0)
|   |   |   |   |   |   |   rs_A > 45: no (5.0)
|   |   rs_PlusMinus > 0
|   |   |   po_A <= 10
|   |   |   |   po_PIM <= 23
|   |   |   |   |   country_group = EURO: no (51.0/4.0)
```

```
|   |   |   |   |   country_group = CAN
|   |   |   |   |   |   rs_GP <= 57: no (12.0)
|   |   |   |   |   |   rs_GP > 57
|   |   |   |   |   |   |   po_A <= 7
|   |   |   |   |   |   |   |   po_A <= 3
|   |   |   |   |   |   |   |   |   Position = D
|   |   |   |   |   |   |   |   |   |   po_PIM <= 2: no (7.0)
|   |   |   |   |   |   |   |   |   |   po_PIM > 2
|   |   |   |   |   |   |   |   |   |   |   po_PIM <= 11: yes (8.0)
|   |   |   |   |   |   |   |   |   |   |   po_PIM > 11: no (4.0/1.0)
|   |   |   |   |   |   |   |   |   Position = C: no (16.0/3.0)
|   |   |   |   |   |   |   |   |   Position = L: no (5.0/1.0)
|   |   |   |   |   |   |   |   |   Position = R
|   |   |   |   |   |   |   |   |   |   Weight <= 205: no (5.0/1.0)
|   |   |   |   |   |   |   |   |   |   Weight > 205: yes (3.0)
|   |   |   |   |   |   |   |   po_A > 3
|   |   |   |   |   |   |   |   |   po_G <= 3: yes (7.0)
|   |   |   |   |   |   |   |   |   po_G > 3
|   |   |   |   |   |   |   |   |   |   po_G <= 5: no (5.0/1.0)
|   |   |   |   |   |   |   |   |   |   po_G > 5: yes (2.0)
|   |   |   |   |   |   |   po_A > 7: no (5.0)
|   |   |   |   |   country_group = USA
|   |   |   |   |   |   CSS_rank <= 36: yes (7.0/1.0)
|   |   |   |   |   |   CSS_rank > 36: no (26.0/3.0)
|   |   |   |   po_PIM > 23
|   |   |   |   |   country_group = EURO: no (2.0)
|   |   |   |   |   country_group = CAN
|   |   |   |   |   |   DraftAge <= 18: yes (7.0)
|   |   |   |   |   |   DraftAge > 18: no (3.0/1.0)
|   |   |   |   |   country_group = USA: yes (4.0/1.0)
|   |   |   po_A > 10: yes (13.0/1.0)
```

=>You can use any tools to implement your ID3 decision tree algorithm

2. Apply the ID3 learner to the hockey draft dataset, using $GP > 0$ as the target class variable. For data preprocessing, drop the *sum_7yr_GP* column.

=>Follow the mentioned preprocessing steps

3. Show the decision tree learned. Which branch is the most informative, meaning that its leaf has the lowest class entropy? Given your understanding of the domain, do the features on the branch make sense?

```
- CSS Ranking plays a vital role in classification.
- CSS_rank<=12, rs_G>15 is the most informative branch
```

4. (Bonus Question) Rerun the Naive Bayes classifier from assignment 1 on the new training and test set for this assignment. Compare the test set accuracy of the decision tree learner to the result of the Naive Bayes classifier.

```
- Decision Tree has better accuracy than Naive Bayes classifier.
As the data gets complex, flexible model like decision tree tends to
perform better than Naive Bayes classifier
```

Theory: Minimum Least Squares Error for Regularized Linear Regression (20 Marks)

Consider least-squares linear regression with L2 regularization as defined in the text.

1. Using the notation of the text, write down the squared-error function, including the regularization term.

$$E(w) = \frac{1}{2}\sum_{1}^{n}(y_n - x_n w)^2 + \frac{\lambda}{2}|w|^2$$

2. Show that the weight vector $\boldsymbol{w}^*$ that minimizes this error function is given by $\boldsymbol{w}^* = (\lambda \boldsymbol{I} + \boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$.

$$\begin{aligned} E(w) =& \frac{1}{2}(y - Xw)^2 + \frac{\lambda}{2}|w|^2 \\ =& \frac{1}{2}(y - Xw)^T(y - Xw) + \frac{\lambda}{2}w^T w \\ =& \frac{1}{2}(y^T y - 2y^T Xw + XwX^T w^T) + \frac{\lambda}{2}w^T w \end{aligned}$$

Taking derivatives on both sides

$$\begin{aligned}\frac{\partial E(w)}{\partial w} =& \frac{\partial(\frac{1}{2}(y^T y - 2y^T Xw + XwX^T w^T))}{\partial w} + \frac{\partial \frac{\lambda}{2} w^T}{\partial w} \\ =& \frac{1}{2} 2X^T Xw - \frac{1}{2} 2X^T y + \lambda w \\ =& X^T Xw - X^T y + \lambda w \\ & X^T Xw + \lambda w = X^T y \\ & (X^T X + \lambda I)w = X^T y\end{aligned}$$

So,

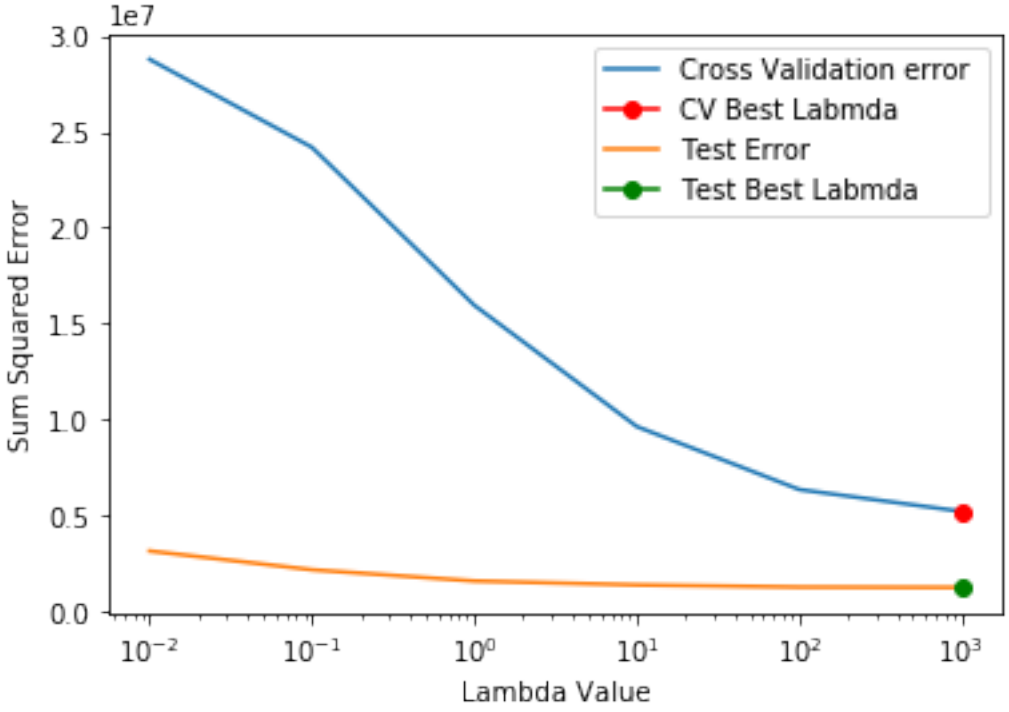
$$w^* = (X^T X + \lambda I)^{-1} X^T y$$

Practice: Implement Least Squares Regression (50 Marks)

1. Try the values from the set $\lambda = 0, 0.01, 0.1, 1, 10, 100, 1000$. Make a plot that shows the following. The horizontal axis shows the value of λ on a log scale (this is called a `semilogx` plot). One curve in the plot should show the squared-error loss evaluated by using 10-fold *cross-validation* on the training set. The second curve shows the squared-error loss evaluated by applying the learned weight vector to the *test set*. Put this plot in your report, and note which regularizer value you would choose from the cross-validation, and which regularizer value would give the lowest squared-error on the test set.

2. For the regularizer that you chose as best from cross-validation, inspect the learned weight magnitudes. Are any of the quadratic interaction terms important (i.e. carry significant weight compared to other variables)? The decision tree also captures interactions among predictor variables - how do the decision tree interactions compare to the interaction terms with high weights?

   `Here is the plot of Least square regression using cross validation and mentioned lambda values`



Note: The value changes between 10^2 to 10^3 for both cross validation and test error based on how you standardize the features ,deal with the 0 value features and select the division of your kfolds. We have considered both these range of lambdas as the correct answer

Figura 1: Least Square regression with Regularization

From the observation of the weight magnitudes in the least square regression model, following are the highest weight predictors with interaction terms responsible for the prediction.
$DraftAge : Weight$
$CSS_rank : rs_G$
$CSS_rank : rs_P$
$Height : Weight$

The interaction terms are similar to that of decision tree splits. This states that the decision tree and the linear regression agrees on the use of interaction terms for the given dataset and using interaction terms can be helpful.