APMA-E4990: Homework 2

**Problem 1. Cross-validation for polynomial regression**

**Root-mean-square error (RMSE) vs Polynomial degree**

A 4th degree polynomial minimizes the RMSE on the testing data.Macintosh HD:Users:brianweinstein:Dropbox:Columbia:4. Spring 2015:Modeling Social Data APMA E4990:modeling-social-data:homework_2:problem_1:rmse_vs_degree.pdf

**Best fit polynomial**

Macintosh HD:Users:brianweinstein:Dropbox:Columbia:4. Spring 2015:Modeling Social Data APMA E4990:modeling-social-data:homework_2:problem_1:best_fit.pdfFull data set with the coefficients and plot of the 4th degree polynomial overlaid.



**Problem 2. Logistic regression for article classification**

**Cross-validation curve**

AUC vs log(Lambda)

Macintosh HD:Users:brianweinstein:Dropbox:Columbia:4. Spring 2015:Modeling Social Data APMA E4990:modeling-social-data:homework_2:problem_2:AUC_vs_logLambda.pdf

**Accuracy and AUC**

Accuracy: 0.905

AUC: 0.9767

**ROC curve for the testing data**

A 4th degree polynomial, indicated with a dashed gray line, minimizes the RMSE on the testing data.

Macintosh HD:Users:brianweinstein:Dropbox:Columbia:4. Spring 2015:Modeling Social Data APMA E4990:modeling-social-data:homework_2:problem_2:ROC_curve.pdf

**Informative words**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Words indicative of a Business article | |  | | Words indicative of a World article | |
| **weight** | **word** |  | **weight** | | **word** |
| 2.266917 | obamas |  | -3.112892 | | burial |
| 1.877835 | collect |  | -2.672924 | | died |
| 1.738303 | nasty |  | -2.655990 | | thailand |
| 1.673156 | updated |  | -2.563773 | | holy |
| 1.396554 | nbc |  | -2.410290 | | anticorruption |
| 1.210733 | company |  | -2.395889 | | canadian |
| 1.178025 | kevin |  | -2.383248 | | abuja |
| 1.126458 | publishing |  | -2.360692 | | videosharing |
| 1.094364 | (Intercept) |  | -2.291570 | | islands |
| 1.081952 | executive |  | -2.245923 | | minister |
| 1.073323 | also |  |  | |  |