**2.874/2.884/10.354/10.554 Homework 4 Process Data Analytics**

Your objective is to construct models for estimating paracetamol concentrations from ATR-FTIR spectra and temperature collected in a batch purification process, continuing from HW3. The original training, validation, and testing datasets are in FTIR\_train.mat, FTIR\_val.mat, and FTIR\_test.mat, respectively, in the file HW-1-data.zip. Each file contains a predictor matrix **X** and the corresponding paracetamol concentration **y**. Each row in the predictor matrix denotes a different observation, while each column denotes absorbance at different frequencies and the last column contains temperature data. The frequencies are contained in the vector freq. Although respective of a common practice in applications, the training, validation, and testing sets were poorly chosen in HW1, as the use of random sampling resulted in highly biased data appearing in all three datasets. Also, HW1 only implemented a single cross-validation, which can lead to inaccurate estimates of the prediction error. In HW3, your tasks were to (a) perform high-quality splits of the complete dataset into training, validation, and test datasets; (b) build models for various sets and estimate the model stability, and (c) apply rigorous cross-validation procedures to estimate the prediction error for both unbiased and biased data.

In HW4, you will apply latent variable methods to build models for high-quality splits of the complete dataset into training, validation, and test datasets and compare the test RMSEs with those obtained by ridge regression, lasso, and elastic net in HW3. As always, training data should be used to train the model parameters, validation data should be used to find optimal hyperparameter(s), and test data should be used to report the average prediction error. You are allowed to include into your analyses any learnings, results, or plots from your HW1 and HW3 solutions and the solution keys provided for those homeworks.

1. For a well-chosen split of the complete dataset into training, validation, and test datasets, apply (i) partial least squares, (ii) principal component regression, and (iii) canonical correlation analysis to build models using the temperature and the absorbance at all frequencies. When validating, plot the RMSE as a function of the number of latent variables for PCA and PLS, and compare to a visualization of the eigenvalues for each component of PCA. Use these visualizations to perform a scree test and determine the optimal number of components for each model. To visualize how these models weight different portions of the spectrum, overlay the regression coefficients on a sample absorbance spectrum for PLS and CCA, and for PCR and PLS, plot an overlay of the first three principal component vectors on a sample absorbance spectrum. Discuss your observations. For all three methods, report the average prediction RMSE for the test data, and discuss your findings. [Note: one way to build the models is to use the Matlab commands pca+regress, plsregress, and canoncorr.]

**Solution: (75 pts)** This solution uses the same split that was chosen in HW1:

|  |  |
| --- | --- |
| **Training Data** | **C = 0.01, 0.02, 0.035** |
| **Validation Data** | **C = 0.025** |
| **Testing Data** | **C = 0.03** |
| **Biased Testing Data** | **C = 0.014** |

Other data splits are valid, as the validation and test datasets are not the outermost concentrations (*C* = 0.01 or *C* = 0.035)

1. **Partial Least Squares:**

A partial least squares regression model was trained on the data using the Matlab plsregress command and z-scored concentrations and centered predictors. Validation was performed by predicting the concentrations for different numbers of model components as shown in Figure a.i.1.

****

**Figure a.i.1:** Validation RMSE vs. model order for partial least squares.

An absolute minimum was found at 8 components, with no gains in predictive power achieved by increasing the model order. [More generally, even if a curve as in Figure a.i.1 did not have a clear minimum and had an RMSE improving slightly by adding additional components, the model should still only use the number of components occurring at the inflection point in order to prevent overfitting.]

The resulting model consists of the coefficients shown in Figure a.i.2.

****

**Figure a.i.2:** Optimal PLS Coefficients vs. an absorbance spectrum.

This model closely matches our physical intuition, heavily weighting the peaks at 1250 cm–1 and 1500 cm–1 while assigning very little importance to the wide absorbance peak near 1650 cm–1 that corresponds to water. Figure a.i.3 examines the components used to construct the regression coefficients.



**Figure a.i.3:** Optimal PLS latent variable coefficients vs. absorbance spectrum. The constant term was omitted in Figure a.i.3 since our concentration data were z-scored.

The first vector weighs the peak at 1550 cm–1 highly, but also places significant weight on noisy data contained between 1600 and 1700 cm–1. The second and third vectors gradually decrease in relative magnitude, approaching smoothed versions of the spectral absorbance plots.

Using these optimal coefficients, we obtain a testing RMSE of **g/g**

Pitfalls:

Did not attempt the question: (−30 pts)

Did not provide testing RMSE (−3 pts)

Missing one or more of the plots outlined in the problem statement (−3 pts/missing)

Performed scree test incorrectly (used a max # of components beyond inflection point) (−2 pts)

Insufficient analysis of findings (−2 pts)

**ii) Principal Component Analysis:**

The training matrix of spectral absorbances and the temperature was split into its principal components using the Matlab pca command. To understand the features extracted by this method, Figure a.ii.1 visualizes the coefficients of the first three principal component vectors against a sample absorbance spectrum. The absorbance spectrum is multiplied by these coefficients to generate the principal components.



**Figure a.ii.1:** The first three principal component coefficients for the spectra.

The first principal component is simply a constant, positive multiplier of the raw data. In components 2 and 3, the coefficient peaks align with relative maxima and minima in the absorption spectra. The principal components were then successively used to train models by linear regression with ordinary least squares. The resulting RMSEs and component eigenvalues were then visualized to determine the optimal model order in Figure a.ii.2.

 

**Figure a.ii.2:** Validation RMSE vs. model order for principal component regression.

The eigenvalue plot and the resulting validation RMSE by each model order show strong agreement that the predictive power is maximized using the first 7 principal components.

This optimized model produces a testing RMSE of  **g/g.**

Pitfalls:

Did not attempt the question: (−30 pts)

Did not provide testing RMSE (−3 pts)

Missing one or more of the plots outlined in the problem statement (−3 pts/missing)

Performed scree test incorrectly (used a max # of components beyond inflection point) (−2 pts)

Insufficient analysis of findings (−2 pts)

1. **Canonical Variance Analysis:**

Canonical component regression was performed using the matlab canoncorr command. Because this decomposition technique produces a unique solution, there was no need for validation. The resulting model is visualized in Figure a.iii.2.



**Figure a.iii.2:** Canonical component regression coefficients compared to absorbance spectra.

This model resulted in a large number of 0-valued coefficients, particularly at low frequencies.

Pitfalls:

Did not attempt the question: (−15 pts)

Did not provide testing RMSE (−3 pts)

Missing the plot requested in the problem statement (−5 pts/missing)

Insufficient analysis of findings (−3 pts)

Did not group data (−3 pts)

**Aside: Cross Validation**

While not required for this assignment, these methods would also have been cross validated in a real application. Cross-validation is difficult to automate for the scree test is difficult to automate due to the subjectivity of assessing when a deviation from linearity occurs. While not best practice, for the sake of simplicity, here we limit the maximum number of PCA and PLS components to 50 each and validate by taking the absolute minimum validation error. Examining the component count for each choice of testing and validation datasets then provides a good indication of which models actively suffer from overfitting and which can use an arbitrarily high number of coefficients but with diminishing returns. If examined manually, they would all have roughly the same number of components (about 6 to 10). Cross-validation of each latent variable method yields the results in Tables a.1 to a.5.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Val:*C*=0.01 | Val:*C*=0.02 | Val:*C*=0.025 | Val:*C*=0.03 | Val:*C*=0.035 |
| Test: *C*=0.01 |  | 5.31×10–5 | 2.12×10–5 | 1.53×10–5 | 2.47×10–5 |
|  | 1.05×10–3 | 7.94×10–4 | 1.02×10–3 | 8.63×10–4 |
| Test: *C*=0.02 | 8.75×10–4 |  | 4.71×10–5 | 2.78×10–5 | 4.09×10–5 |
| 1.39×10–4 |  | 1.71×10–4 | 1.67×10–4 | 9.28×10–5 |
| Test: *C*=0.025 | 6.65×10–4 | 1.10×10–4 |  | 2.0×10–5 | 1.56×10–5 |
| 6.21×10–5 | **5.86**×10–5 |  | 3.78×10–5 | 3.91×10–5 |
| Test: *C*=0.03 | 3.53×10–4 | 1.96×10–5 | 2.02×10–5 |  | 4.17×10–5 |
| 6.18×10–5 | 5.15×10–5 | 4.35×10–5 |  | 3.72×10–5 |
| Test: *C*=0.035 | 6.60×10–4 | 2.21×10–5 | 2.22×10–5 | 3.72×10–5 |  |
| 9.0×10–5 | 1.39×10–4 | 3.47×10–5 | 4.17×10–5 |  |

**Table a.1:** Partial least squares validation and unbiased testing root-mean-squared errors in units of g/g. The values in yellow are validation errors, and in blue are unbiased test errors.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Val:*C*=0.01 | Val:*C*=0.02 | Val:*C*=0.025 | Val:*C*=0.03 | Val:*C*=0.035 |
| Test: *C*=0.01 |  | 6 | 10 | 6 | 7 |
| Test: *C*=0.02 | 4 |  | 6 | 6 | 6 |
| Test: *C*=0.025 | 4 | 9 |  | 9 | 10 |
| Test: *C*=0.03 | 10 | 16 | 8 |  | 5 |
| Test: *C*=0.035 | 10 | 7 | 6 | 5 |  |

**Table a.2:** The number of components used in the model minimizing the validation error for partial least squares

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Val:*C*=0.01 | Val:*C*=0.02 | Val:*C*=0.025 | Val:*C*=0.03 | Val:*C*=0.035 |
| Test: *C*=0.01 |  | 3.13×10–5 | 2.08×10–5 | 1.53×10–5 | 1.9×10–5 |
|  | 1.07×10–3 | 8.67×10–4 | 9.08×10–4 | 8.85×10–4 |
| Test: *C*=0.02 | 8.84×10–4 |  | 5.54×10–5 | 3.44×10–5 | 1.71×10–5 |
| 1.26×10–4 |  | 1.10×10–4 | 2.03×10–4 | 1.75×10–4 |
| Test: *C*=0.025 | 3.24×10–4 | 1.10×10–4 |  | 2.3×10–5 | 1.47×10–5 |
| 1.06×10–4 | 5.54×10–5 |  | 2.87×10–5 | 3.23×10–5 |
| Test: *C*=0.03 | 4.19×10–4 | 1.21×10–5 | 2.09×10–5 |  | 4.0×10–5 |
| 4.38×10–5 | 4.54×10–5 | 3.88×10–5 |  | 3.1×10–5 |
| Test: *C*=0.035 | 6.60×10–4 | 1.65×10–5 | 2.24×10–5 | 3.1×10–5 |  |
| 4.8×10–5 | 7.65×10–5 | 7.36×10–5 | 4.0×10–5 |  |

**T****able a.3:** Principal component regression validation and unbiased testing root-mean-squared errors in units of g/g. The values in yellow are validation errors, and in blue are unbiased test errors.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Val:*C*=0.01 | Val:*C*=0.02 | Val:*C*=0.025 | Val:*C*=0.03 | Val:*C*=0.035 |
| Test: *C*=0.01 |  | 6 | 35 | 24 | 35 |
| Test: *C*=0.02 | 48 |  | 30 | 6 | 6 |
| Test: *C*=0.025 | 4 | 30 |  | 38 | 17 |
| Test: *C*=0.03 | 50 | 44 | 22 |  | 6 |
| Test: *C*=0.035 | 50 | 21 | 8 | 6 |  |

**Table a.4:** The number of components used in the model minimizing the validation error for partial least squares.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Val:*C*=0.01 | Val:*C*=0.02 | Val:*C*=0.025 | Val:*C*=0.03 | Val:*C*=0.035 |
| Test: *C*=0.01 |  | 1.55×10–3 | 9.13×10–4 | 6.58×10–4 | 1.10×10–3 |
| Test: *C*=0.02 | 4.94×10–4 |  | 2.23×10–4 | 2.81×10–4 | 1.46×10–4 |
| Test: *C*=0.025 | 4.45×10–5 | 6.790×10–5 |  | 6.62×10–5 | 4.65×10–5 |
| Test: *C*=0.03 | 7.70×10–5 | 1.62×10–4 | 5.11×10–5 |  | 2.11×10–4 |
| Test: *C*=0.035 | 1.63×10–4 | 8.93×10–5 | 1.33×10–4 | 8.21×10–4 |  |

**Table a.5:** Canonical variance regression unbiased testing root-mean-squared errors in units of g/g.

1. Compare the results from the latent variable methods in part a with the results from ridge regression, lasso, and elastic net in HW1. Following your mathematical intuition, explain any observed patterns in the testing RMSEs, and compare the coefficient plots generated for PLS and CCA to those of the models built previously with regularized regression.

**Solution (25 points)**:

The obtained testing RMSEs for this dataset are

|  |  |  |
| --- | --- | --- |
| PLS | PCR | CCA |
| 4.37×10–5 | 3.84×10–5 | 5.11×10–5 |
| Ridge | Lasso | Elastic Net |
| 2.25×10–5 | 5.12×10–5 | 4.83×10–5 |

Quantitatively, the obtained testing RMSEs are all the same order of magnitude, suggesting similar predictive power. To gain a better qualitative understanding of the relative model quality, we examine the coefficient plots in Figures b.1 and b.2.

****

**Figure b.1:** Latent variable coefficients vs. an absorbance spectrum.



**Figure b.2**: Regularized regression coefficients vs. an absorbance spectrum.

From our domain knowledge, we know that the wide peak around 1600 to 1700 cm–1 corresponds to water and offers poorer predictive power compared to the rest of the dataset. Therefore, a strong model should place a great deal of weight on the peak at 1550 cm–1 and assign comparatively little weight to the absorbance peaks corresponding to water, which is accomplished very well by the regularized regression methods, particularly lasso and elastic net. It is worth noting that PLS bears striking resemblance to the model generated by ridge regression, but with an even higher relative weighting assigned to the peak near 1550 cm–1.

The canonical correlation coefficients are very different from the other models, assigning no weight at all to the peaks between 1200 cm–1 and 1300 cm–1 and seemingly randomly assigning a massive portion of the correlation between spectrum absorbance and concentration to about 1420 cm–1. In general, this method is not recommended for comparing a matrix of inputs to a single output vector, and is much more valuable when used to examine the relationships between multiple highly correlated inputs and outputs.

For principal component analysis, looking at the coefficients used to construct the first three components, we note that the first component is a simple integration of the spectrum, whereas progressively higher order coefficients begin to selectively amplify portions of the spectrum. The second and third components assign nearly equal weighting to the water peak and the peak of interest near 1550 cm–1, suggesting that PLS and the regularized regression models did a better job of cutting through noise in the spectral data.

Pitfalls:

Did not attempt the question: (−25 pts)

Did not compare testing RMSEs (−10 pts)

Did not provide graphical analysis (−10 pts)

Provided graphical analysis, but used different data splits between the models   
in question (−3 pts)