cs109a_hw5_209

October 23, 2018

1 CS109A Introduction to Data Science:

1.1 Homework 5 AC 209 : PCA

Harvard University Fall 2018 Instructors: Pavlos Protopapas, Kevin Rader
Names of people you have worked with goes here:

```
In [1]: # RUN THIS CELL FOR FORMAT
        import requests
        from IPython.core.display import HTML
        styles = requests.get("https://raw.githubusercontent.com/Harvard-IACS/2018-CS109A/maste
        HTML(styles)

Out[1]: <IPython.core.display.HTML object>

In [2]: # Imports
        import numpy as np
        import matplotlib.pyplot as plt
        from sklearn.datasets import make_regression
        from sklearn.linear_model import LinearRegression, Ridge, Lasso, ElasticNet, RidgeCV, I
        from sklearn.model_selection import train_test_split
        from sklearn.metrics import r2_score

        %matplotlib inline
```

Question 5 [25 pts]

Suppose we want to conduct PCA on the model matrix $X \in \Re^{np}$, where the columns have been suitably set to zero mean. In this question, we consider the squared reconstruction error:

$$\| XQ - XQ_m \|^2$$

for a suitable set of eigenvectors forming the matrix Q_m, as discussed below. Suppose that we conduct eigendecomposition of $X^T X$ and obtain eigenvalues $\lambda_1, \ldots, \lambda_p$ and principal components Q, i.e.

$$X^T X = Q \Lambda Q^T$$

5.1 Suppose that the matrix norm is simply the squared dot product, namely

$$\| A \|^2 = A^T A$$