École polytechnique
MAP 534: introduction to machine learning
A. Durmus alain.durmus@polytechnique.edu

2 Linear model for regression

Exercise 2.1. Let $\mathsf{F} \subset \mathbb{R}^n$ be a linear subspace of $\mathbb{R}^n$. Consider $\{\mathbf{e}_i\}_{i=1}^d$ a orthonormal basis of F.

(1) Show that

$$\mathrm{proj}_{\mathsf{F}} : x \in \mathbb{R}^n \mapsto \sum_{i=1}^{d} \langle x, \mathbf{e}_i \rangle \, \mathbf{e}_i \ , \tag{1}$$

is a linear map from $\mathbb{R}^n \to \mathsf{F}$.

(2) Give its matrix representation.

(3) Show that for any $x \in \mathbb{R}^n$, the problem

$$\text{minimize } \|x - y\| \text{ over } y \in \mathsf{F} \ , \tag{2}$$

has a unique solution given by $\mathrm{proj}_{\mathsf{F}}(x)$.

We have shown that

Proposition 2.1. *Let $\mathsf{F} \subset \mathbb{R}^n$ be a linear subspace of $\mathbb{R}^n$. Consider $\{\mathbf{e}_i\}_{i=1}^d$ a orthonormal basis of F. Then, the problem*

$$\textit{minimize } \|x - y\| \textit{ over } y \in \mathsf{F} \ , \tag{3}$$

has a unique solution given by $\mathrm{proj}_{\mathsf{F}}(x)$,

$$\mathrm{proj}_{\mathsf{F}} : x \in \mathbb{R}^n \mapsto \sum_{i=1}^{d} \langle x, \mathbf{e}_i \rangle \, \mathbf{e}_i \ . \tag{4}$$

Exercise 2.2. Recall the Gram-Schmidt orthogonal theorem:

Theorem 2.2. *Let $\mathsf{F} \subset \mathbb{R}^n$ be a linear space and $\{\mathbf{f}_i\}_{i=1}^d$ a basis of F. There exists an orthonormal basis $\{\mathbf{e}_i\}_{i=1}^d$ of F such that for any $j \in \{1, \dots, d\}$, $\mathrm{span}(\mathbf{f}_1, \dots, \mathbf{f}_j) = \mathrm{span}(\mathbf{e}_1, \dots, \mathbf{e}_j)$.*

Show how to construct recursively $\{\mathbf{e}_i\}_{i=1}^d$.

Exercise 2.3. The purpose of this exercise is to show the following result:

Theorem 2.3 (Singular value decomposition)**.** *Let $\mathbf{A} \in \mathbb{R}^{n \times m}$. There exist two orthogonal matrices $\mathbf{U} \in \mathbb{R}^{n \times n}$ and $\mathbf{V} \in \mathbb{R}^{m \times m}$, $\Sigma \in \mathbb{R}^{n \times m}$ such that $\Sigma_{i,j} = 0$ for $i \neq j$ and $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^t$.*

Show this result in the particular case where $n = m$.

2.1 Exercises

Exercise 2.4. Consider some data $\{(y_i, x_i)\}_{i=1}^n$ such that $y_i \in \mathbb{R}$ and $x_i = (x_i^{(1)}, \dots, x_i^{(d)}) \in \mathbb{R}^d$. We consider linear regression models for these and consider the error function

$$E(w) = \frac{1}{2}\sum_{i=1}^{n}(f_w(x_i) - y_i)^2 = \|\mathbf{X}w - y\|^2/2 \,, \quad \text{since } f_w(x_i) = x_i^{\mathrm{T}} w \,, \tag{5}$$

where

$$y = (y_1, \dots, y_n)^{\mathrm{T}} \in \mathbb{R}^n \,, w = (w_1, \dots, w_d)^{\mathrm{T}} \in \mathbb{R}^d \,, \tag{6}$$

$$\mathbf{X} = (x_1, \dots, x_n)^{\mathrm{T}} = \begin{pmatrix} x_1^{(1)} & \dots & x_1^{(d)} \\ . & . & . \\ . & . & . \\ . & . & . \\ x_n^{(1)} & \dots & x_n^{(d)} \end{pmatrix} \in \mathbb{R}^{n \times d} \,, \tag{7}$$

and the least squares estimator as follows:

$$\hat{w} \in \operatorname{argmin} E(w) \,.$$

1. Show that the estimator $\hat{w}$ is always well defined for all $\{(y_i, x_i)\}_{i=1}^n$.
2. Show that if $d \leq n$, then $w \mapsto \mathbf{X}w$ is injective if and only if $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is invertible.
3. We now suppose that $d \leq n$. Show that if $w \mapsto \mathbf{X}w$ is injective then the estimator of least squares is unique and given for all $\{(y_i, x_i)\}_{i=1}^n$ by

$$\hat{w} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \,. \tag{8}$$

Exercise 2.5. Sometimes, computing directly the inverse $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ may be numerically instable using Gauss Jordan elimination. A solution to this problem is to use the pseudo-inverse of this matrix. The pseudo-inverse of a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ is defined as a matrix $\mathbf{A}^\dagger \in \mathbb{R}^{m \times n}$ such that

$$\mathbf{A}\mathbf{A}^\dagger\mathbf{A} = \mathbf{A} \,, \qquad \mathbf{A}^\dagger\mathbf{A}\mathbf{A}^\dagger = \mathbf{A}^\dagger \,. \tag{9}$$

(1) Show that if $\mathbf{A}$ is invertible (so $n = m$), any pseudo-inverse equals $\mathbf{A}^{-1}$.

(2) Using a singular value decomposition for $\mathbf{A}$, show how to compute in practice a pseudo-inverse of $\mathbf{A}$.

(3) Deduce another method to compute the inverse of $\mathbf{A}$ using a singular value decomposition for $\mathbf{A}$.

2.2 Homework

In all your homeworks, you can not use any library except numpy, seaborn, pandas and matplotlib.

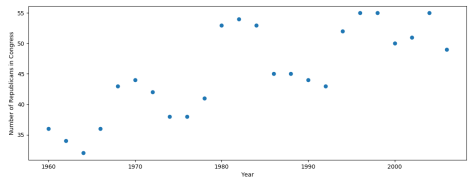
Exercise 2.6 (Homework)**.** The objective of this problem is to learn about linear regression with basis functions by modeling the number of Republicans in the Senate. The file

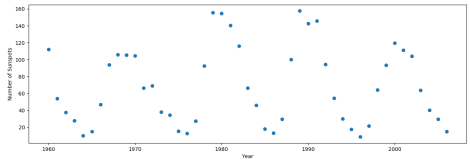
`year-sunspots-republicans.csv`

contains the data you will use for this problem. It has three columns. The first one is an integer that indicates the year. The second is the number of Sunspots observed in that year. The third is the number of Republicans in the Senate for that year. The data file looks like this:

```
Year,Sunspot_Count,Republican_Count
1960,112.3,36
1962,37.6,34
1964,10.2,32
1966,47.0,36
```

You can see scatterplots of the data in the figures below. The horizontal axis is the Year, and the vertical axis is the Number of Republicans and the Number of Sunspots, respectively.





(Data Source: http://www.realclimate.org/data/senators_sunspots.txt)

In this problem you need to implement least squares regression using 4 different basis functions for **Year (x-axis)** v. **Number of Republicans in the Senate (y-axis)**.

1. Load the dataset and plot figures similar to the ones above.

The numbers in the *Year* column are large (between 1960 and 2006), especially when raised to various powers. To avoid numerical instability due to ill-conditioned matrices in most numerical computing systems, we will scale the data first: specifically, we will scale all "year" inputs by subtracting 1960 and then dividing by 40.

2. Implement these procedures to obtain new features. In the sequel, we only use these new features.

3. Plot the data and regression lines for each of the following sets of basis functions, and include the generated plot as an image in your submission. You will therefore make 4 total plots:

   (a) $\phi_j(x) = x^j$ for $j = 1, \ldots, 5$
   ie, use basis $y = a_1x^1 + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5$ for some constants $\{a_1, ..., a_5\}$.

   (b) $\phi_j(x) = \exp -(40x - \mu_j)^2/25$ for $\mu_j = 0, 5, 10, \ldots, 50$

   (c) $\phi_j(x) = \cos(x/j)$ for $j = 1, \ldots, 5$

   (d) $\phi_j(x) = \cos(x/j)$ for $j = 1, \ldots, 25$

   * Note: Please make sure to add a bias term for all your basis functions above in your implementation

4. For each plot include the train error.

Exercise 2.7 (homework)**.** The Boston Housing Dataset The Boston Housing Dataset is a derived from information collected by the U.S. Census Service concerning housing in the area of Boston MA. The following describes the dataset columns:

- CRIM - per capita crime rate by town

- ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS - proportion of non-retail business acres per town.
- CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- NOX - nitric oxides concentration (parts per 10 million)
- RM - average number of rooms per dwelling
- AGE - proportion of owner-occupied units built prior to 1940
- DIS - weighted distances to five Boston employment centres
- RAD - index of accessibility to radial highways
- TAX - full-value property-tax rate per $\$10,000$
- PTRATIO - pupil-teacher ratio by town
- B - $1000(Bk - 0.63)^2$ where Bk
- LSTAT - % lower status of the population
- MEDV - Median value of owner-occupied homes in \$1000 's

We will build a regression model to predict CRIM using the other variables INDUS, NOX, RM, RAD, TAX, PTRATIO, LSTAT and MEDV as features. The file

`housing.csv`

contains the data you will use for this problem.

(1) For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

(2) Fit a multiple regression model to predict the response using all of the predictors. Describe your results.

(3) How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.

(4) Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor X, fit a model of the form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$$

(5) If y denotes the list containing the CRIM variable observations and x is the features matrix, apply the following transformation:

```
y =  np.log1p(y)
for col in x.columns:
if np.abs(x[col].skew()) > 0.3:
x[col] = np.log1p(x[col])
```

Then, fit a multiple regression model to predict the response using all of the predictors. Describe your results. Compare the resulting error loss with the one previously obtained without any transformations.