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

2 Linear model for regression

Exercise 2.1. Consider some data $\{(y_i, x_i)\}_{i=1}^n$ such that $y_i \in \mathbb{R}$ and $x_i = (x_i^{(1)}, \ldots, x_i^{(d)}) \in \mathbb{R}^d$. We consider linear regression models for these and consider the error function

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

where

$$y = (y_1, \ldots, y_n)^{\mathrm{T}} \in \mathbb{R}^n \;, w = (w_1, \ldots, w_d)^{\mathrm{T}} \in \mathbb{R}^d \;, \tag{2}$$

$$\mathbf{X} = (x_1, \ldots, x_n)^{\mathrm{T}} = \begin{pmatrix} x_1^{(1)} & \ldots & x_1^{(d)} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ x_n^{(1)} & \ldots & x_n^{(d)} \end{pmatrix} \in \mathbb{R}^{n \times d} \;, \tag{3}$$

and the least squares estimator as follows:

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

1. Show that the estimator 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

$$\hat{w} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \;. \tag{4}$$

We suppose now that $d \leq n$, $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is invertible and that

$$y_i = w^{\mathrm{T}} x_i + \sigma \epsilon_i \;, \tag{5}$$

where $w \in \mathbb{R}^d$ and $\epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, 1)$. We call the vector of residuals, the vector

$$\hat{\epsilon} = y - \mathbf{X}\hat{w} = [\mathrm{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}]y \;. \tag{6}$$

4. Show that:

$$\text{(i) } \hat{w} \sim \mathrm{N}(w, \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}) \;\text{ and }\; \text{(ii) } \hat{\epsilon} \sim \mathrm{N}(0, \sigma^2(\mathrm{I}_n - \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}})) \;.$$

5. Deduce an unbiased estimator of w and σ^2. An estimator $\tilde{w}$ of w is said to be linear if there exists a matrix $\mathbf{A}$ such that

$$\tilde{w} = \mathbf{A}y \;.$$

   Moreover, an estimator $\tilde{w}$ is said to be unbiased if $\mathbb{E}[\tilde{w}] = w$.

6. Show that for any unbiased linear estimator $\tilde{w}_{\mathbf{A}}$ associated to the matrix $\mathbf{A}$, there exists a positive symmetric matrix $\mathbf{R}$ such that that $\mathrm{Cov}(\tilde{w}_{\mathbf{A}}) = \mathrm{Cov}(\hat{w}) + \mathbf{R}$.

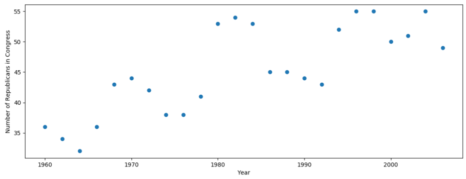
Exercise 2.2 (Homework)**.** The objective of this problem is to learn about linear regression with basis functions by modeling the number of Republicans in the Senate. The file

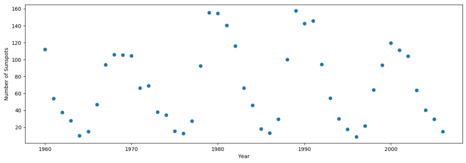
`year-sunspots-republicans.csv`

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 $y = a_1 x^1 + a_2 x^2 + a_3 x^3 + a_4 x^4 + a_5 x^5$ 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.

5. Repeat the same exact process as above but for **Number of Sunspots (x-axis)** v. **Number of Republicans in the Senate (y-axis)**. Here, to avoid numerical instability with numbers in the *Sunspot_Count* column, we will also scale the data first by dividing all "sunspot count" inputs by 20. In addition, only use data from before 1985, and only use basis functions (a), (c), and (d) – ignore basis (b). You will therefore make 3 total plots. For each plot make sure to also include the train error.

6. Which of the three bases (a, c, d) provided the "best" fit? **Choose one**, and keep in mind the generalizability of the model.

7. Given the quality of this fit, do you believe that the number of sunspots controls the number of Republicans in the senate (Yes or No)?