Introduction to Machine Learning (67577)

Exercise 2 Linear Regression

Second Semester, 2023

Student Name: Yosef Edery Anahory.

ID: 345175475

IML - Ex2 - Theoretical part

Let $X_{m \times n}$ and let $y \in \mathbb{R}^m$

2.1 Solutions of the normal equation:

1. prove: $\ker(X) = \ker(X^T X)$

$\ker(X) = \{ \alpha \in \mathbb{R}^n \mid X \cdot \alpha = 0^m \}$

- Notice that for all $0 \neq \alpha \in \ker(X)$ $\quad X\alpha = 0 \Rightarrow X^T \cdot (X\alpha) = 0$ (trivial)
- Let $\alpha \in \ker(X^T X)$ then $\quad X^T X \cdot \alpha = 0$ (by kernel definition)
  Notice that $X^T \cdot X \cdot \alpha = \alpha^T \cdot X^T \cdot X \cdot \alpha = (X\alpha)^T \cdot (X\alpha) = \|X\alpha\|^2 \Rightarrow X\alpha = 0 \Rightarrow \alpha \in \ker(X)$
  $\{ X^T \cdot X \cdot \alpha = 0 \}$ ■

2. Prove $\mathrm{Im}(A^T) = \ker(A)^\perp$ s.t $A_{n \times n}$

Let's prove $\mathrm{Im}(A) = \ker(A^T)^\perp$ (which is the same since $(A^T)^T = A$)
Let $y \in \mathrm{Im}(A)$. From image definition we get
that there exists $\alpha \in \mathbb{R}^n$ s.t $A \cdot \alpha = y$.

Let $\beta \in \ker(A^T)$, $A^T \cdot \beta = 0 \Rightarrow \langle A^T \beta, \alpha \rangle = \langle \beta, A^T \alpha \rangle = \langle \beta, x \rangle = \langle x, \beta \rangle$
From the orthogonal complement definition we get:
Then, $\mathrm{Im}(A) \subseteq \ker(A^T)^\perp$.

Let's prove $\ker(A^T)^\perp \subseteq \mathrm{Im}(A)$:

To do this we'll get $y \notin \mathrm{Im}(A)$ and prove $y \in \ker(A^T)^\perp$. It's enough to find $\alpha \in \ker(A^T)$ s.t $\langle y, \alpha \rangle \neq 0$. Since we assumed $y \notin \mathrm{Im}(A)$, the vector α need to have a component in $\mathrm{Im}(A)^\perp$.
Let $\alpha \in \mathrm{Im}(A)^\perp$ s.t $\langle b, c \rangle \neq 0$.
Since $\alpha \in \mathrm{Im}(A)^T$ we get that α is orthogonal to any vector in $\mathrm{Im}(A)$. Indeed, $\langle c, AA^T c \rangle = 0$.
Then:

$$\|A^T \alpha\|^2 = \langle A^T \alpha, A^T \alpha \rangle = \langle \alpha, AA^T \alpha \rangle = 0$$

Thus: $A^T \alpha = 0 \Rightarrow \alpha \in \ker(A^T)$ s.t α is as wanted.

3. Let $y = Xw$ prove: the system has ∞ solutions $\Leftrightarrow y \perp \ker(X^T)$
   X is not invertible therefore it has ∞ or infinite solutions or no solution.
   The system has at least one solution if and only if $y \in \mathrm{Im}(X)$.

   Then we get from question 2 and this last lemma that the system has
   infinite solutions if and only if $y \in \ker(X^T)^\perp \Leftrightarrow y \perp \ker(X^T)$
   ($\perp$ definition) ■

4. Let $X^T X w = X^T y$. Prove that normal equations can only have 1 solution (if $X^T X$ is invertible) or ∞ solutions (otherwise).
   Proof: Let's start by $X^T X$ not being invertible. We have proved $\ker(X^T X) = \ker(X)$ and
   $X^T y \perp \ker(X^T X) \Leftrightarrow$ system has ∞ solutions $\Rightarrow$ Enough to prove $X^T y \perp \ker(X)$. N.t:

   $\forall \beta \in \ker(X) \quad \langle \beta, X^T y \rangle = \langle X\beta, y \rangle = 0$ ■ All in all, the system has 1 solution (if $X^T X$ is invertible) or

∞ solutions.

5.

a) Show that P is symmetric:

Proof: Sum of symmetric matrices is symmetric, therefore P is symmetric.

b) Prove eigenvalues of P are 0 or 1 and $v_1, \dots, v_k$ are the eigenvectors:

$$Pv_j = \sum_{i=1}^{k} v_i v_i^T v_j = \sum_{i=1}^{k} v_i \delta_{ij} = v_j$$

c) Show that $\forall v \in V \; Pv = v$.

$$x \in V \Rightarrow x = \sum_{i=1}^{k} \alpha_i v_i \Rightarrow Px = P\sum_{i=1}^{k} \alpha_i v_i = \sum_{i=1}^{k} \alpha_i P v_i = \sum_{i=1}^{k} \alpha_i v_i$$

d) Prove that $P^2 = P$

$$P^2 = UDU^T UDU^T = UDDU^T = UDU^T = P$$

e) Prove that $(I-P)D = 0$

$$(I-P)\cdot P = P - P^2 = P - P = 0$$

2.3. Least squares

6. Show that $X^T X$ is invertible

Let's substitute X with X's SVD. Let $X = U\Sigma V^T$ be X's SVD then:

$$[X^T X]^{-1} X^T = [(U\Sigma V^T)^T (U\Sigma V^T)]^{-1} (U\Sigma V^T)^T =$$
$$= [V\Sigma^T \Sigma V^T]^{-1} V\Sigma^T U^T =$$
$$= V(\Sigma^T\Sigma)^{-1} V^T V \Sigma^T U^T = VD^{-1}\Sigma^T U^T$$

s.t $D := \Sigma^T\Sigma$ and $D_{(\text{rank}(X)\times\text{rank}(X))}$ s.t it is a diagonal matrix with $D_{ii} = \sigma_i^2$. W.l.o.g as the columns of X are linearly independent then $\sigma_1 \geq \dots \geq \sigma_{d-1} > 0$ and then D^{-1} is also defined.
W.l.o.g:

$$[D^{-1}\Sigma^T]_{ii} = \frac{1}{\sigma_i^2}\sigma_i = \frac{1}{\sigma_i} = \Sigma^+_{ii}$$

Then we conclude all in all that:

$$(X^T X)^{-1} X^T y = V\Sigma^+ U^T y = X^+ y \quad \square$$

7. Show that $X^T X$ is invertible if and only if $\{x_1, \dots, x_m\} = \mathbb{R}^d$

Proof: X's rank is equal subspaces's dimension $\{x_1, \dots, x_m\}$.
By the SVD the theorem implies that X's rank is equal to the rank $X^T X$.
W.l.o.g $x_1, \dots, x_m$ spans $\mathbb{R}^d$ iff $(X^T X) = d$. Since $X^T X$ is a $d \times d$ matrix, it is invertible iff its rank is d.

8. Let $X = U\Sigma V^T$ be the SVD decomposition of X. Let r be the rank of X, and recall

$$V = [V_1\ V_2] \quad U = [U_1\ U_2] \quad \Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix}$$

s.t. V_1, U_1 represent first (r) columns and V_2, U_2 the rest of the columns of U, V.
Σ_1 is the diagonal matrix.

Given w, let's define $b = U^T w$ and $b_1 = U_1^T w$, $b_2 = U_2^T w$ and therefore $b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$. U is orthonormal matrix, which means is isometry (then $\|Ux\| = \|x\|$ for any vector x) then we get $\|w\| = \|b\|$, so b is the minimal norm.

V is also isometry then:

$$\|y - X^T w\|^2 = \|y - V\Sigma U^T w\|^2 = \|V(V^T y - \Sigma b)\|^2 = \|V^T y - \Sigma b\|^2 = \left\| [V_1, V_2]^T y - \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \right\|^2 =$$

$$= \|V_1^T y - \Sigma_1 b_1\|^2 + \|V_2^T y\|^2$$

$\|\cdot\|$ is all, to get a minimal solution for $\|y - Xw\|$:

$$\Sigma_1 b_1 - V_1^T y = 0 \iff \Sigma_1 b_1 = V_1^T y \iff b_1 = \Sigma_1^{-1} V_1^T y.$$

In order to minimize b:

$$b_1 = \Sigma_1^{-1} V_1^T y \quad \text{and} \quad b_2 = 0.$$

W.t. $\hat{w} = X^{T\dagger} y$ we get:

$$U_1^T \hat{w} = U_1^T X^{T\dagger} y = U_1^T U_1 \Sigma_1^{-1} V_1^T y = \Sigma_1^{-1} V_1^T y$$

and from:

$$U_2^T \hat{w} = U_2^T X^{T\dagger} y = U_1^T U_1 \Sigma_1^{-1} V_1^T y = 0$$

For any other solution $\bar{w}$, we get that the first condition needs to be satisfied, but the second may be not. Then: $\|\hat{w}\| \le \|\bar{w}\|$. ∎

3. Practical Part

3.1 Linear Regression

Preprocess data implementation:

1. Which features to keep and which not?
2. Which features are categorical how how did you treat them?
3. What other features did you design and what is the logic behind creating them?
4. How did you treat invalid/missing values?
5. Explain any additional processing performed on the data

1) The features that I chose to keep, were the ones that represent relevant data that can influent the price by its value such us bathrooms or sqft_living. The higher number of bathrooms or sqft_living implies a higher price of the house. On the other hand: id, lat or long don't provide consistent information, therefore I've decided to remove them.

2) Some of the categorical features are zipcode. Which it has an importance on the price since is a good parameter to determine if is well located. Otherwise, a higher zipcode doesn't implies a higher house sell price. Therefore, to profit this valuable data I've added number of zipcode column and assigned a binary value to each of them. One other categorical feature is yr_renovated. This feature is also valuable since is measuring the quality of the place. But the problem that those houses who were not renovated are receiving a value of 0 (even if they were built recently). Therefore, I've changed the columns to "recently renovated" which is a binary value. To calculate the value, I've picked the top 20% renovated years to be 1 and others 0.

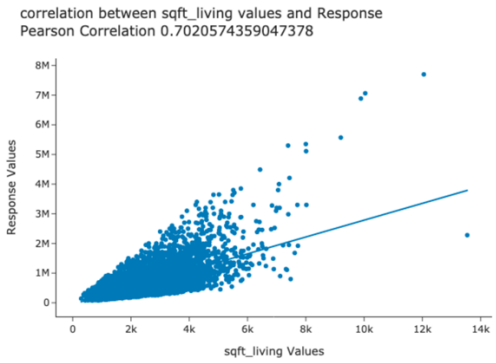
3) One other feature that I've design is the five_years_period_build_(year) which divides the built year on a period of five year classifying the data with a binary value.

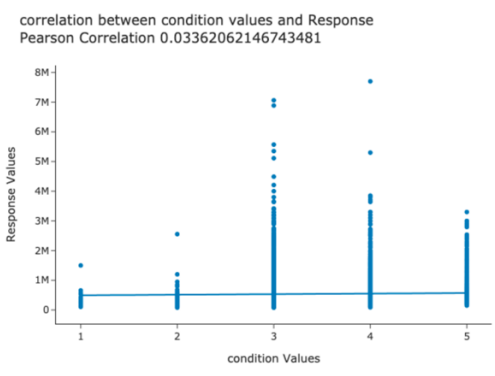
4) Invalid missing values were removed from the training samples and once data was cleaned the average of each future from the training samples was calculated. Then once we preprocess the training data for every NULL value that appear on either of the feature columns we replace it by the average already calculated from the training samples.

5) No other additional processing was performed.

Feature evaluation

Choose two features, one that seems to be beneficial for the model and one that does not:



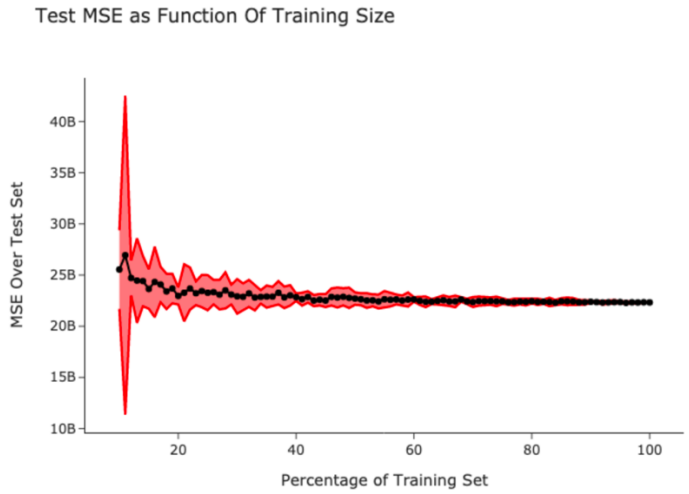


- Notice that the sqft living has high pearson correlation which indicates a good relation sqft_living-price. On the secong graph we can see that the pearson correlation equals 0.03 which is a low value and indicates now real impact of the condition values on the price.

Linear Regression fitting

What can we learn about the estimator yˆi in terms of estimator properties?

- We can learn that as the training data increases over the test, the average and variance of loss over test set decreases.



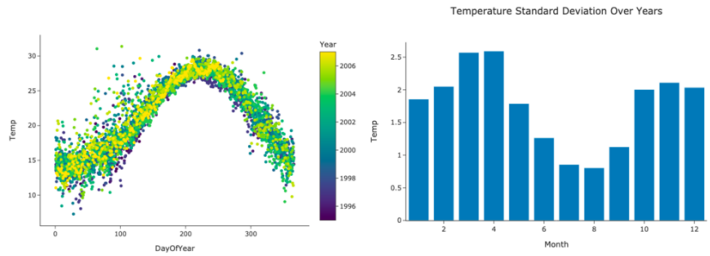
3.2 Polynomial Fitting

What polynomial degree might be suitable for this data?

- Looking at the first graph, we can see that the data displays a consistent undulating pattern over the years, with the highest temperatures appearing around days ±200. It's important to point out that the temperature stabilizes after peaking instead of continuing to decline on both sides. This observation suggests that the data may require a higher-degree polynomial than 2 for an accurate model.

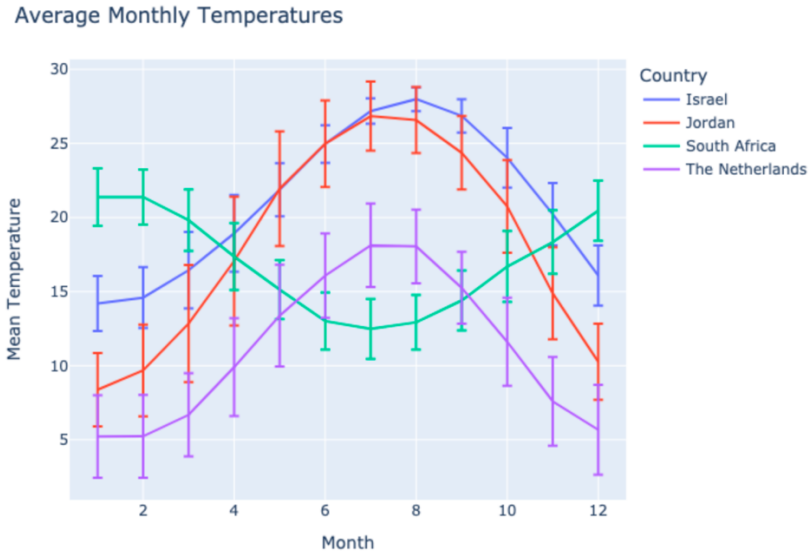
Based on this graph, do you expect a model to succeed equally over all months or are there times of the year where it will perform better than on others?

- The model may perform better during months with low variability (June-September) and worse during months with high variability (March-April). It will perform better since we are measuring variability, low variability implies a more accuracy fitting.
-

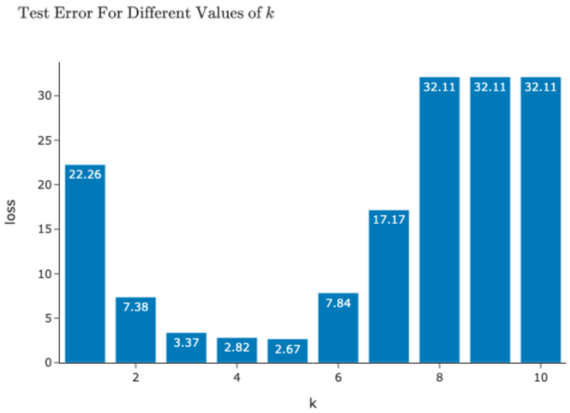


3. Based on this graph, do all countries share a similar pattern? For which other countries is the model fitted for Israel likely to work well and for which not?

- It's worth noting that if we use a model trained on Israeli data to predict temperatures, we'll find that Jordan, which is near Israel, displays a similar pattern to Israel, with a small deviation. On the other hand, when we analyze data from Holland, we can achieve accurate results by reducing the average temperature by approximately 10 degrees. However, South Africa behaves in the exact opposite way to Israel, with warm temperatures when Israel is cold and vice versa. Therefore, we'll need to train a model on data from a different region.

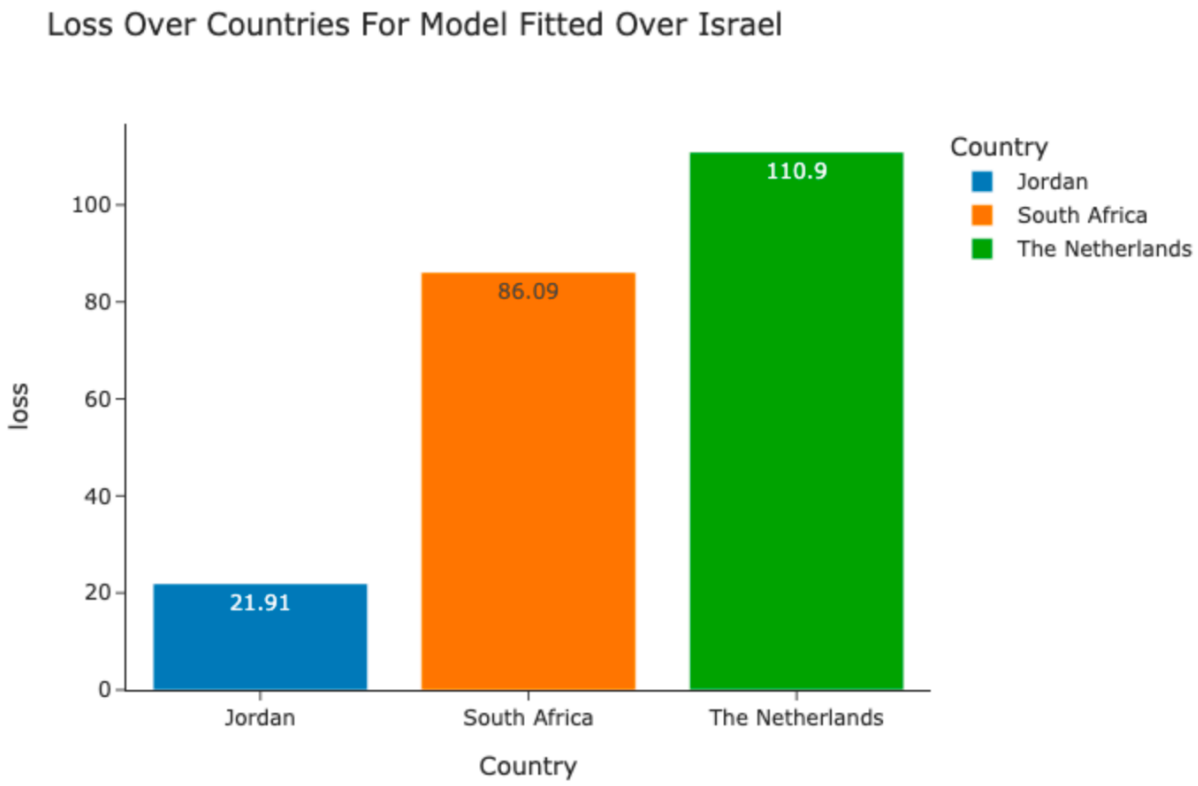


4. Print the test error recorded for each value of k. In addition plot a bar plot showing the test error recorded for each value of k. Based on these which value of k best fits the data? In the case of multiple values of k achieving the same loss select the simplest model of them. Are there any other values that could be considered?



- Notice that when k=5 the loss is the minimum, hence a 5th degree polynomial model will bring the prediction closer to being the best.

5. Fit a model over the entire subset of records from Israel using the k chosen above. Plot a bar plot showing the model's error over each of the other countries. Explain your results based on this plot and the results seen in question 3.



- After fitting a model to a subset of observations from Israel, it was discovered that this model was not as effective when applied to data from other countries. However, the distribution of temperatures in Jordan closely resembled that of Israel. Consequently, out of the three countries examined, the model performed most accurately on Jordan. On the other hand, the temperature distributions in The Netherlands and South Africa were less similar to that of Israel, resulting in a poorer fit for the model.