Introduction to Machine Learning (67577)

Exercise 2 - Linear Regression

Dor Roter
208772251

April 28, 2021

1 Theoretical Questions

1.1 Solutions of the Normal Equations

1. Let $X \in \mathbb{R}^{m \times d}$ be a matrix, then

$$Ker(X) = \left\{ v \in \mathbb{R}^d \,|Xv = 0 \right\} = \left\{ v \in \mathbb{R}^d \,\middle|X^T Xv = X^T \cdot 0 \right\} = \left\{ v \in \mathbb{R}^d \,\middle|X^T Xv = 0 \right\} = Ker(X^T X) \quad \square$$

2. Let $A \in \mathbb{R}^{n \times n}$.
firstly, let us show that for all $v \in Im(A^T)$ it holds that $v \in Ker(A)^\perp$:

$$\begin{aligned} & v \in Im(A^T) \Rightarrow v = A^T x \\ \Rightarrow & \forall u \in Ker(A) \qquad \langle v|u \rangle = v^T \cdot u = \left(A^T x\right)^T \cdot u = x^T A \cdot u = x^T \cdot Au = x^T \cdot 0 = 0 \\ \text{Therefore: } & v \in Im(A^T) \Rightarrow \langle v|x \rangle = 0 \quad \forall x \in Ker(A) \Leftrightarrow v \in Ker(A)^\perp \end{aligned}$$

as such, it holds that $Im(A^T) \subseteq Ker(A)^\perp$.

let us show that for all $v \in Ker(A)^\perp$ it holds that $v \in Im(A^T)$.
now, we assume by contradiction that $v \notin Im(A^T)$ and we will show that $v \notin Ker(A)^\perp$.

$$\begin{aligned} & v \notin Im(A^T) \Rightarrow v \in Im(A^T)^\perp \\ therefore: \quad & \forall u \in \mathbb{R}^n \quad v^T \cdot \left(A^T u\right) = (Av) \cdot u = 0 \end{aligned}$$

as such, it holds that Av must be 0, which means $v \in Ker(A)$.
but, $v \in Ker(A) \Rightarrow v \notin Ker(A)^\perp$.
Therefore, it holds that $Im(A^T) \supseteq Ker(A)^\perp$.

$\Rightarrow Im(A^T) = Ker(A)^\perp \; \square$

3. Let $y = Xw$ be a non-homogeneous system of linear equations, with X as square not invertible matrix.
As X is not invertible, the equations system could not have a single solution, Therefore there are either 0, or infinitly many solutions for the provided linear equations set.

There might be a solution to the equations iff $y \in Im(X)$ (could be writen as a linear combination of X using the scalars specified in w).
Therefore:

$$y \in Im(X) \underset{q.2}{=} Ker(X^T)^{\perp}$$

$\Rightarrow y \perp Ker(X^T)$ $\square$

4. Let $X^T X w = X^T y$ be a normal linear system.
We will address the two following possibilities for $X^T X$ seprately.

   (a) $X^T X$ **is invertible:** then $X^T X w = X^T y \Leftrightarrow w = \left(X^T X\right)^{-1} X^T y$ and we have found a single unique solution for the linear system.

   (b) $X^T X$ **is not invertible:** since $X^T X$ is a square matrix, by the lemma proved in q.3:

   $$X^T y \perp Ker(\left(X^T X\right)^T) = Ker(X^T X) \Leftrightarrow \text{"there are } \infty \text{ solutions for } X^T X w = X^T y\text{"}$$

   Let us show $X^T y \perp Ker(X^T X)$:

   $$\forall v \in Ker(X^T X) \quad \left\langle X^T y | v \right\rangle = \left(X^T y\right)^T v = y^T X v$$

   since we have proved in $q.1$ that $Ker\left(X\right) = Ker(X^T X)$, it holds that $v \in Ker(X^T X) \Leftrightarrow v \in Ker(X) \Leftrightarrow Xv = 0$ and thus:

   $$\underbrace{\forall v \in Ker(X^T X) \quad \left\langle X^T y | v \right\rangle = y^T X v = y^T \cdot 0 = 0}_{\Updownarrow \; X^T y \perp Ker(X^T X)}$$

1.2 Projection Matrices

5. Let $V \subseteq \mathbb{R}^d, dim(V) = k$, $v_1, \ldots, v_k$ be orthonormal basis of V, and $P = \sum_{l.=1}^{k} v_l v_l^T$.

   (a) by P's defenition $P_{i,j} = \sum_{l=1}^{k} \left(v_l\right)_i \cdot \left(v_l\right)_j = \sum_{l=1}^{k} \left(v_l\right)_j \cdot \left(v_l\right)_i = P_{j,i}$ and therefore P's symetric.

   (b) Since P is a symetric matrix, it holds that P has an EVD decomposition $P = UDU^T$ where U is an orthogonal matrix and D is a diagonal matrix.
   Therefore, $P^2 = UDU^T \cdot UDU^T = UD^2U^T$.
   By the claim proved in (d), $P^2 = P$, and so $UD^2U^T = P^2 = P = UDU^T$.
   As $diag(\lambda_1^2, \ldots, \lambda_k^2) = D^2 = D = diag(\lambda_1, \ldots, \lambda_k)$, it holds that for each $i \in [k]$ $\lambda_i = \lambda_i^2 \Leftrightarrow \lambda_i \in \{0, 1\}$.
   Finally for all $i \in [k]$ it holds that:

   $$Pv_i = \left(\sum_{l=1}^{k} v_l v_l^T\right) \cdot v_i = \sum_{l=1}^{k} \left(v_l v_l^T \cdot v_i\right) = \sum_{l=1}^{k} v_l \left(v_l^T \cdot v_i\right) = \sum_{l=1}^{k} v_l \left\langle v_l | v_i \right\rangle \underbrace{=}_{v_1,\ldots,v_k \text{ are orthonormal}} \sum_{l=1}^{k} v_l \cdot \delta_{i,l} = v_i$$

   $$\left\langle v_l | v_i \right\rangle = \begin{cases} 1 & l = i \\ 0 & l \neq i \end{cases}$$

   and therefore $v_1, \ldots, v_k$ are the the eigenvectors corresponding to the eigenvalue 1.

   (c) Let $v \in V$, therefore there is a linear combination of the orthonormal basis of V, $v_1, ..., v_k$ and $\alpha_1, ..., \alpha_k \in \mathbb{R}$ such as $v = \sum_{l=1}^{k} \alpha_i \cdot v_i$.

Therefore:

$$Pv = \left(\sum_{i=1}^{k} v_i v_i^T\right)\left(\sum_{j=1}^{k} \alpha_j v_j\right) = \sum_{i=1}^{k}\sum_{j=1}^{k}\left(v_i v_i^T \cdot \alpha_j v_j\right) = \sum_{i=1}^{k}\sum_{j=1}^{k}\alpha_j v_i \langle v_i | v_j\rangle = \sum_{i=1}^{k}\sum_{j=1}^{k}\alpha_j v_i \cdot \delta_{i,j} = \sum_{i=1}^{k}\alpha_i v_i = v$$

(d) Using P's defenition it holds that:

$$P^2 = \left(\sum_{i=1}^{k} v_i v_i^T\right)\cdot\left(\sum_{j=1}^{k} v_j v_j^T\right) = \sum_{i=1}^{k}\sum_{j=1}^{k} v_i v_i^T v_j v_j^T = \sum_{i=1}^{k}\sum_{j=1}^{k} v_i \langle v_i | v_j\rangle v_j^T = \sum_{i=1}^{k}\sum_{j=1}^{k} v_i \cdot \delta_{i,j} \cdot v_j^T = \sum_{i=1}^{k} v_i \cdot v_i^T = P$$

(e) Using the pervious lemmas:

$$(I-P)P = P - P^2 \underset{(d)}{=} P - P = 0$$

1.3 Least Squares

6. Let $X = U\Sigma V^T$ be the SVD decomposition of X.
Firstly, we will show $\left(X^T X\right)^{-1} = V\left(\Sigma^T\Sigma\right)^{-1}V^T$:

$$X^T X = V\Sigma^T U^T \cdot U\Sigma V^T = V\Sigma^T\Sigma V^T$$

$$\Rightarrow \begin{cases} X^T X \cdot V\left(\Sigma^T\Sigma\right)^{-1}V^T = V\Sigma^T\Sigma\underbrace{V^T\cdot V}_{I}\left(\Sigma^T\Sigma\right)^{-1}V^T = V\underbrace{\left(\Sigma^T\Sigma\right)\cdot\left(\Sigma^T\Sigma\right)^{-1}}_{I}V^T = V\cdot V^T = I \\ V\left(\Sigma^T\Sigma\right)^{-1}V^T\cdot X^T X = V\left(\Sigma^T\Sigma\right)^{-1}\underbrace{V^T\cdot V}_{I}\Sigma^T\Sigma V^T = V\underbrace{\left(\Sigma^T\Sigma\right)^{-1}\cdot\left(\Sigma^T\Sigma\right)}_{I}V^T = V\cdot V^T = I \end{cases}$$

And so

$$\hat{w} = \left(X^T X\right)^{-1}X^T y = V\left(\Sigma^T\Sigma\right)^{-1}\underbrace{V^T\cdot V}_{I}\Sigma^T U^T\cdot y = V\left(\Sigma^T\Sigma\right)^{-1}\Sigma^T U^T\cdot y$$

We will demostrate now that $\left(\Sigma^T\Sigma\right)^{-1}\Sigma^T = \Sigma^\dagger$ and by so prove the general solution from the recitation equals to the one seen in class.
Since Σ is a diagonal matirx, $\left(\Sigma^T\Sigma\right)^{-1}_{i,j} = \begin{cases}\sigma_i^{-2} & \sigma_i \neq 0 \\ 0 & \sigma_i = 0\end{cases}$ and so, $\left(\left(\Sigma^T\Sigma\right)^{-1}\Sigma^T\right)_{i,j} = \begin{cases}\sigma_i^{-2}\sigma_i = \sigma_i^{-1} & \sigma_i \neq 0 \\ 0\cdot\sigma_i = 0 & \sigma_i = 0\end{cases} =$
$\Sigma^\dagger$.
Therefore, $\hat{w} = \left(X^T X\right)^{-1}X^T y = V\left(\Sigma^T\Sigma\right)^{-1}\Sigma^T U^T\cdot y = V\Sigma^\dagger U^T\cdot y = X^\dagger y$ □

7. let $X \in \mathbb{R}^{m\times d}$,then $T_X : \mathbb{R}^d \to \mathbb{R}^m$ is the respective linear transformation represented by X.
for $X^T X \in \mathbb{R}^{d\times d}$ it holds that:

$$X^T X \in \mathbb{R}^{d\times d} \text{ is invertible} \Leftrightarrow Ker\left(X^T X\right) = 0 \underset{q.1}{\Leftrightarrow} Ker(X) = 0 \Leftrightarrow \underbrace{rank(X) = dim(\mathbb{R}^d)}_{\text{rank–nullity theorem}} - Ker(X) = d$$

And therefore:

$$\begin{aligned} & rank(X) = dim\left(Span\left(x_1, \ldots, x_m\right)\right) = d \\ \Leftrightarrow & Span\left(x_1, \ldots, x_m\right) = \mathbb{R}^d \end{aligned}$$

As stated $X^T X \in \mathbb{R}^{d \times d}$ is invertible $\Leftrightarrow Span\left(x_1, \ldots, x_m\right) = \mathbb{R}^d$. $\square$

8. Let us define $\hat{w} = X^{\dagger}y$, and let $\overline{w} \in \mathbb{R}^d$ be another solution of the linear equations set $X^T X w = X^T y$.

   We have seen in recitation that $X = U\Sigma V^T = \begin{bmatrix} U_{\mathcal{R}} & U_{\mathcal{N}} \end{bmatrix} \begin{bmatrix} \mathcal{S} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_{\mathcal{R}}^T \\ V_{\mathcal{N}}^T \end{bmatrix} = U_{\mathcal{R}} \mathcal{S} V_{\mathcal{R}}^T = \tilde{U}\tilde{\Sigma}\tilde{V}^T$ where $\mathcal{S}$ is a diagonal invertible $r \times r$ matrix.

   Therefore, each w that solves $X^T X w = X^T y$ must be of form $w = \tilde{V}\tilde{\Sigma}^{-1}\hat{U}^T y$ where $\tilde{V}\tilde{\Sigma}^{-1}\hat{U}^T$ could each be padded into the original SVD decomposition dimentions.

   Let us note that under this notation, $\hat{w} = V\Sigma^{\dagger}U^T y$ and $\overline{w} = \begin{bmatrix} V_{\mathcal{R}} & V_{\mathcal{N}} \end{bmatrix} \begin{bmatrix} \mathcal{S}^{-1} & & & & 0 \\ & \sigma_{r+1} & & & 0 \\ & & \ddots & & 0 \\ & & & \sigma_d & 0 \end{bmatrix} \begin{bmatrix} U_{\mathcal{R}}^T \\ U_{\mathcal{N}}^T \end{bmatrix}$,

   therefore also $\forall i \in [r]\ \hat{w}_i = \overline{w}_i$ and it holds that:

$$||\overline{w}||^2 \underset{\text{pythegorean theorem}}{=} \sum_{i=1}^{d} \overline{w}_i^2 = \sum_{i=1}^{r} \overline{w}_i^2 + \sum_{i=r+1}^{d} \overline{w}_i^2 \geq \sum_{i=1}^{r} \overline{w}_i^2 + \sum_{i=r+1}^{d} 0 = \sum_{i=1}^{r} \hat{w}_i^2 + \sum_{i=r+1}^{d} 0 = ||\hat{w}||^2$$

   $\Rightarrow ||\hat{w}|| \leq ||\overline{w}||$ for each $\overline{w}$ solution of $X^T X w = X^T y$. $\square$

2 Practical Questions

13. First, we can note the categorical features of the dataset include the following features: $waterfront, view, condition, grade, z$ while $waterfront, view, condition, grade$ are simple boolean values or ordinal values, they are already already encoded as a needed for us, thus it is left to us to decide on the proper handling of $zipcode, lat$ and $long$.
    In the preprocessing process I have also clustered some non-categorical features, such as yr_built and $yr\_renovated$, and transformed them into categorical dummy values in order to allow the model to better deal with theirs illinearity.
    Next, I have also broken up "$sqft\_basement$" into another binary column stating wether a basement exists, and using the passed time between the last renovation/building I've stated in "is_new" wether the building is relatively new (providing the linear model with the ability to take into account the combination of "yr_build" and "$yr\_renovated$").
    Lastly, in order to allow the model to "understand" geolocation, I have tupled a rough estimation of the latitude and longtitude data into a single column, "$location$" which is in and of itself a categorical feature of an area (roughly 10km in diameter).
    I have then encoded those new "$zipcode$","$location$","yr_built","$yr\_renovated$" features using one-hot encoding.

14. Scree-plot of non-categoric design matrix:

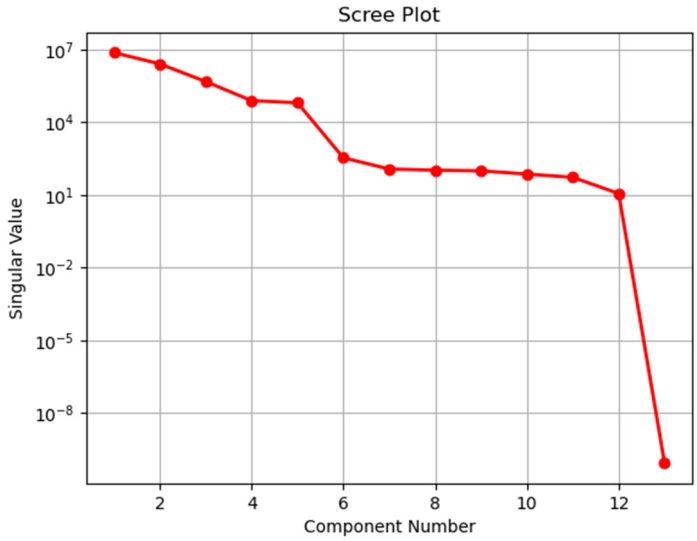

Figure 1: Scree-Plot of the design matrix

15. We note the data seems to have some close to linearliy dependent features since part of the singular values are close to zero. This indicated the model would probably function well using a subset of the features available to us, and might generalize better if we do so (using just 5-10 features).
    Furthermore, the high singular values are corresponding to a singular-vector which, as we have proven, is one of the features vectors, and thus those values correspond to important features for the model's prediciton.

16. Plotting the training progress as a factor of p - the percent of the training dataset used we note that the MSE tends to be higher, and less stable when a smaller portion of the dataset is used to fit the model, and as p's value surpasses 20% we can see the MSE settling towards it's minimum point, with only minur improvments to the models accuracy.

Therefore, we can conclude that fitting this model requires a minimum of around 3.5 thousand samples in order to porducde a somewhat credible estimator.

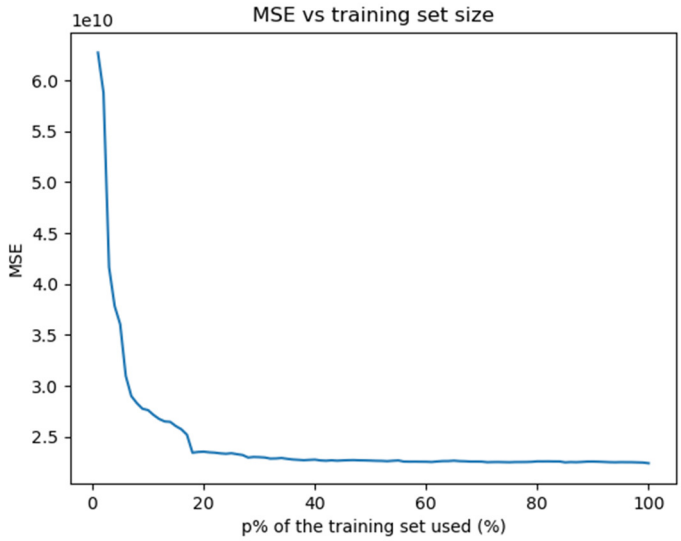

Figure 2: MSE plot by training set size

17. We notice by looking at the correlation coefficent of each feature in relation to the response vector that "$sqft\_living$" and "$grade$" columns, with respective coefficents of 0.68, and 0.66, seem very benefical to the model as their linear correlation with the response vector, described by their pearson correlation is close to 1, thus both rise and fall in unison with the response vector. This hipotesis is backed by the plots of both features against the response vector showing a rather clear linear relation.

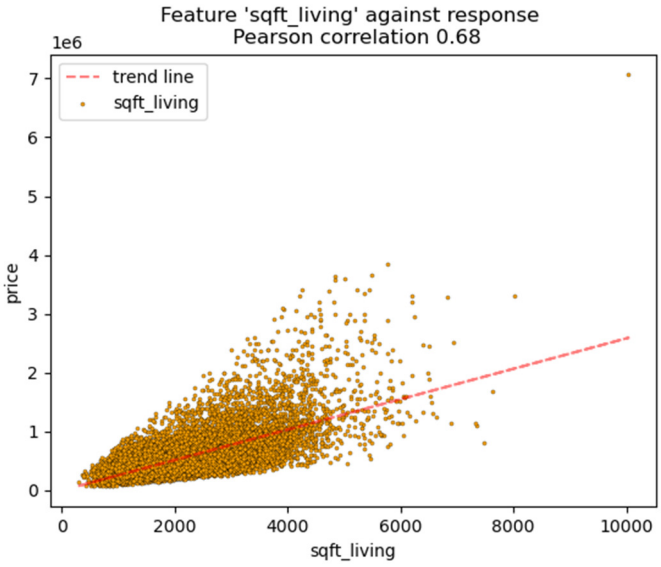

Figure 3: "$sqft\_living$" vs "$price$"

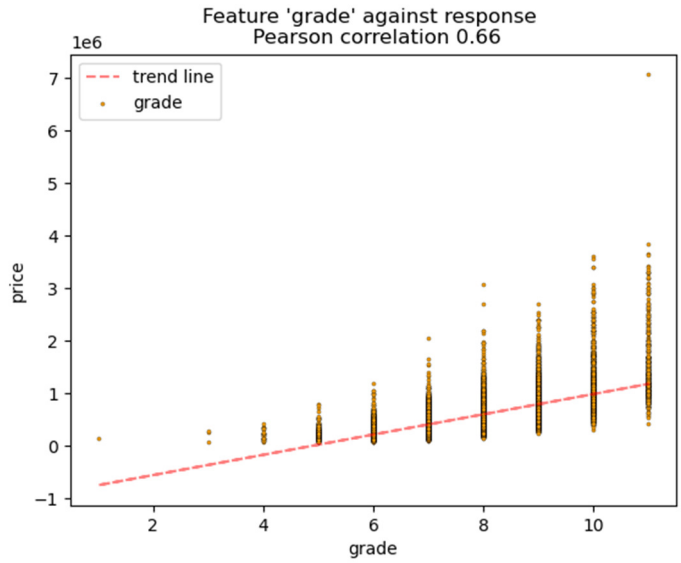

Figure 4: "*grade*" vs "*price*"

As opposed to these columns, the "*condition*" column has a pearson correlation coefficient of only 0.05, thus indicating there's not much of a linear relation ship between the "*condition*" column and the response vector, and as such, this column seems less beneficial for the linear regression model's prediction

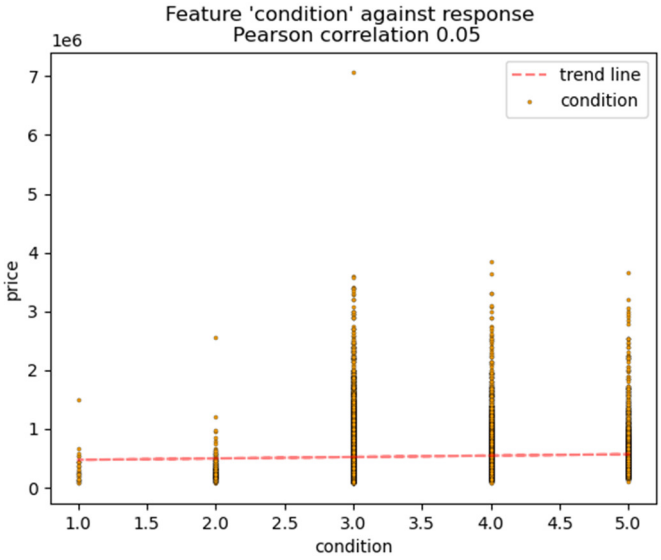

Figure 5: "*condition*" vs "*price*"