Introduction to Machine Learning (67577)
Exercise II – Mathematical Background
Hadar Sharvit – 208287599

Theoretical Questions

$X \in \mathbb{R}^{m\times d}$ – m rows (samples) & d column (features)

$y \in \mathbb{R}^m$ – response vector corresponding to samples in X.

Solutions of the Normal Equations

1. Prove that $\mathrm{Ker}(X) = \mathrm{Ker}(X^TX)$

   In the first direction, let $v \in \mathrm{Ker}(X)$, therefore $Xv = 0$. Multiply both sides with X^T we have $X^TXv = 0$ which means by definition $v \in \mathrm{Ker}(X^TX)$
   In the other direction, let $v \in \mathrm{Ker}(X^TX)$, therefore $X^TXv = 0$. Multiply both sides by v^T we have $v^TX^TXv = 0 \to (Xv)^TXv = 0 \to ||Xv||^2 = 0$. Since $||Xv||^2 \geq 0$ we have $||Xv||^2 = 0 \Leftrightarrow Xv = 0 \to x \in \mathrm{Ker}(X)$

2. Let $A \in \mathbb{R}^{n\times n}$, prove that $\mathrm{Im}(A^T) = \mathrm{Ker}(A)^\perp$.
   In the first direction, let $v \in \mathrm{Im}(A^T)$. If so, $\exists x \in \mathbb{R}^n$ s.t $A^Tx = v$. We wish to show that $v \in \mathrm{Ker}(A)^\perp$
   Let us consider some arbitrary $w \in \mathrm{Ker}(A)$. From the definition of the kernel, such w satisfies $Aw = 0$. Now we calculate $\langle v, w\rangle$:
   $$\langle v, w\rangle = \langle A^Tx, w\rangle = (A^Tx)^Tw = x^TAw = x^T0 = 0$$
   From here we conclude that v and w are orthogonal, therefore $v \in \mathrm{Ker}(A)^\perp$
   In the second direction, let $v \in \mathrm{Ker}(A)^\perp$. We wish to show that $v \in \mathrm{Im}(A^T)$. This is the same as showing that $v \notin \mathrm{Im}(A^T) \Rightarrow v \notin \mathrm{Ker}(A)^\perp$. In other words, assuming $v \notin Im(A^T)$, we must show that there exists some vector $c \in \mathrm{Ker}(A)$ s.t $\langle v, c\rangle \neq 0$. From here we will conclude that $v \notin \mathrm{Ker(A)}^\perp$, because for v to be in $\mathrm{Ker(A)}^\perp$ it must satisfy $\langle v, u\rangle = 0\ \forall u \in \mathrm{Ker}(A)$, yet we have found one vector $u = c$ for which the claim will not hold. If so, let us find such c:
   Assuming $v \notin \mathrm{Im(A^T)}$, it must have some component$\in \mathrm{Im}(A^T)^\perp$. Let c be that component, i.e., we can choose $c \in \mathrm{Im(A^T)}^\perp$ (note that for such c we have $\langle v, c\rangle \neq 0$ because $v \notin \mathrm{Im}(A^T)$). This c is orthogonal to any vector in $\mathrm{Im(A^T)}$
   $$||Ac||^2 = \langle Ac, Ac\rangle = (Ac)^TAc = c^TA^TAc = \langle c, A^TAc\rangle \overset{(*)}{=} 0$$
   If so, $Ac = 0 \to c \in \mathrm{Ker}(A)$
   $(*)$ For a vector $x = Ac \in \mathbb{R}^n$ we have that $A^Tx \in \mathrm{Im(A^T)}$, and since c is orthogonal to any vector in $\mathrm{Im(A^T)}$, it is orthogonal to $A^Tx = A^TAc$, therefore the product is 0

3. Let $y = Xw$ where $X \in \mathbb{R}^{n\times n}$ non-invertible. Show that the system has ∞ solutions $\Leftrightarrow y \perp \mathrm{Ker}(X^T)$
   $$y \perp \mathrm{Ker}(X^T) \Leftrightarrow y \in \mathrm{Ker}(X^T)^\perp \overset{from\ 2}{\Longleftrightarrow} y \in \mathrm{Im}(X) \Leftrightarrow \exists w \quad s.t\ y = Xw$$
   Now, since X is invertible than $y = Xw$ has either 0 or ∞ solutions, yet since there is at least one solution, there are ∞. Such condition can be rewritten as
   $$y \in \mathrm{Im}(X) \Leftrightarrow y = Xw \text{ has } \infty \text{ solutions}$$
   Which finishes the proof.

4. Prove that the normal linear system $X^TXw = X^Ty$ can only have unique solution (if X^TX is invertible) or infinitely many solutions (otherwise):
   For the case where X^TX is invertible we can write a unique solution for w as followed

$$w = (X^TX)^{-1}X^Ty$$

For the case where X^TX is not invertible

$$\text{the system has } \infty \text{ solutions}$$
$$[\text{from q. 3}] \Leftrightarrow X^Ty \perp \mathrm{Ker}\big(X^TX\big)$$
$$[\text{from q. 1}] \Leftrightarrow X^Ty \perp \mathrm{Ker}(X)$$

Indeed, for some $u \in \mathrm{Ker}(X)$ we have $\langle X^Ty, u\rangle = (X^Ty)^Tu = y^TXu \overset{Xu=0}{=} 0$, therefore $X^Ty \in \mathrm{Ker}(X)^\perp \Rightarrow$ the system has ∞ solutions

Projection Matrices

$V \subseteq \mathbb{R}^d$, $\{v_1, \dots, v_k\}$ = orthogonal basis of V. $P = \sum_{i=1}^k v_iv_i^T$

5. A. P is symmetric ($P = P^T$):

$$P^T = \left(\sum_{i=1}^k v_iv_i^T\right)^T \overset{(*)}{=} \sum_{i=1}^k (v_iv_i^T)^T = \sum_{i=1}^k (v_i^T)^Tv_i^T = \sum_{i=1}^k v_iv_i^T = P$$

where $(*)$ is due to the fact that the transpose of a sum is the sum of transposes

B. the eigenvalues of P are $\lambda = 0{,}1$ and $v_1, \dots, v_k$ are the eigenvectors corresponding to $\lambda = 1$:
We can use the fact that $P^2 = P$ (proof in 5.D) as followed: the eigenvalues and vectors associated with P satisfy the equation

$$Pv = \lambda v$$

Therefore, we can multiply both sides with P and get

$$P^2v = P\lambda v = \lambda Pv = \lambda^2 v$$

Now using $P^2 = P$

$$Pv = \lambda^2 v$$

Combining this with the above equation we get

$$\lambda v = \lambda^2 v$$
$$v(\lambda - \lambda^2) = 0$$
$$\lambda = 0{,}1$$

In section 5.C we see that $\forall v \in V$ $Pv = 1v$, therefore for $v \in \{v_1, \dots, v_k\}$ the claim also holds. That is, $v_1, \dots v_k$ are the eigenvectors corresponding to the eigenvalue $\lambda = 1$

C. $\forall v \in V$ $Pv = v$:
If $v \in V$ then we can write v as a linear combination of $\{v_1, \dots, v_k\}$: $v = \sum_{i=1}^k c_iv_i$ for some $c_i \in \mathbb{R}$.
Therefore,

$$Pv = \sum_{i=1}^k v_iv_i^T \sum_{i=1}^k c_iv_i = \sum_{i,j}^k v_iv_i^Tc_jv_j = \sum_{i,j}^k c_jv_iv_i^Tv_j = \sum_{i,j}^k c_jv_i\langle v_i, v_j\rangle = \sum_{i,j}^k c_jv_i\delta_{ij} = \sum_{i=1}^k c_iv_i = v$$

D. $P^2 = P$:
For $v \in V$ we can use section C:

$$Pv = v \to PPv = Pv \to P^2v = Pv \to P^2 = P$$

Otherwise, we must generalize for all $v \notin V$ (which also holds for $v \in V$):

$$P^2 = P \cdot P = \sum_{i=1}^k v_iv_i^T \sum_{i=1}^k v_iv_i^T = \sum_{i,j=1}^k v_iv_i^Tv_jv_j^T = \sum_{i,j}^k v_iv_i^Tv_jv_j^T = \sum_{i,j}^k v_i\langle v_i, v_j\rangle v_j^T$$

Since $\{v_i\}$ is an orthonormal basis we get $\langle v_i, v_j\rangle = \delta_{ij}$, so we can re-write P^2 as

$$= \sum_{i,j}^k v_i\delta_{ij}v_j^T = \sum_{i=1}^k v_iv_i^T = P$$

E. $(1 - P)P = 0$:

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

Least Squares

6. Show that if X^TX is invertible, $(X^TX)^{-1}X^T = X^\dagger = V\Sigma^\dagger U^T$:

$$\begin{gathered}(X^TX)^{-1}X^T = \\ [X = U\Sigma V^T]\text{:}\ \ ((U\Sigma V^T)^T U\Sigma V^T)^{-1}(U\Sigma V^T)^T = \\ [\text{Transposing}]\text{:}\ \ (V\Sigma^T\Sigma V^T)^{-1}V\Sigma^T U^T = \\ [\text{Inverting}]:\ = (V^T)^{-1}\Sigma^{-1}(\Sigma^T)^{-1}V^{-1}V\Sigma^TU^T = \\ [V^{-1}V = I]\text{:}\ \ (V^T)^{-1}\Sigma^{-1}(\Sigma^T)^{-1}\Sigma^TU^T = \\ [(\Sigma^\text{T})^{-1}\Sigma^T = I]:\ = (V^T)^{-1}\Sigma^{-1}U^T = \\ [V^T = V^{-1}]\text{:}\ \ V\Sigma^{-1}U^T = V\Sigma^\dagger U^T\end{gathered}$$

7. Show that X^TX is invertible iff $\text{span}\{x_1, \dots x_m\} = \mathbb{R}^d$:

   We know from definition that $\text{rank}(X) = \dim\big(span(\{x_1, \dots, x_m\})\big)$, and in class we saw that $\text{rank}(X) = \text{rank}(X^TX)$. For the set $\{x_1, \dots x_m\}$ to span $\mathbb{R}^d$ it must be of dimension d, which in turn means that $\text{span}\{x_1, \dots x_m\} = \mathbb{R}^d \Leftrightarrow \dim\big(span(\{x_1, \dots, x_m\})\big) = d \Leftrightarrow \text{rank}(X^TX) = d$. From here we can use the fact that X^TX has dimension $d \times d$, therefore it is invertible by definition iff $\text{rank}(X^TX) = d$

   Simply written, we have

$$\begin{gathered}\text{span}\{x_1, \dots x_m\} = \mathbb{R}^d \\ \Leftrightarrow \dim\big(span(\{x_1, \dots, x_m\})\big) = d \\ \Leftrightarrow \text{rank}(X) = d \\ \Leftrightarrow \text{rank}(X^TX) = d \\ \Leftrightarrow X^TX \text{ is invertible}\end{gathered}$$

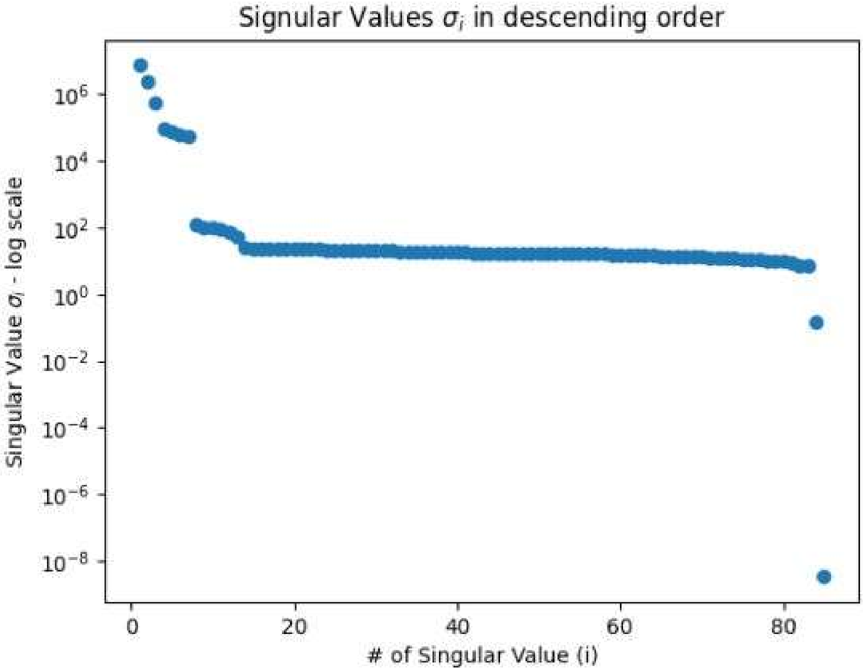
8. show that for any other solution $\overline{w}$, $||\hat{w}|| \leq ||\overline{w}||$, where $\hat{w} = X^\dagger y$ (this is the case where X^TX is not invertible):

   we know that for $i \in \{1, \dots, r\}$ $\overline{w}_i = \hat{w}_i$, and for $i \in \{r+1, \dots d\}$ $\hat{w}_i$ = 0, and $\overline{w}_i$ could by any value $\geq$ 0. This means that the ℓ_2 norm satisfies

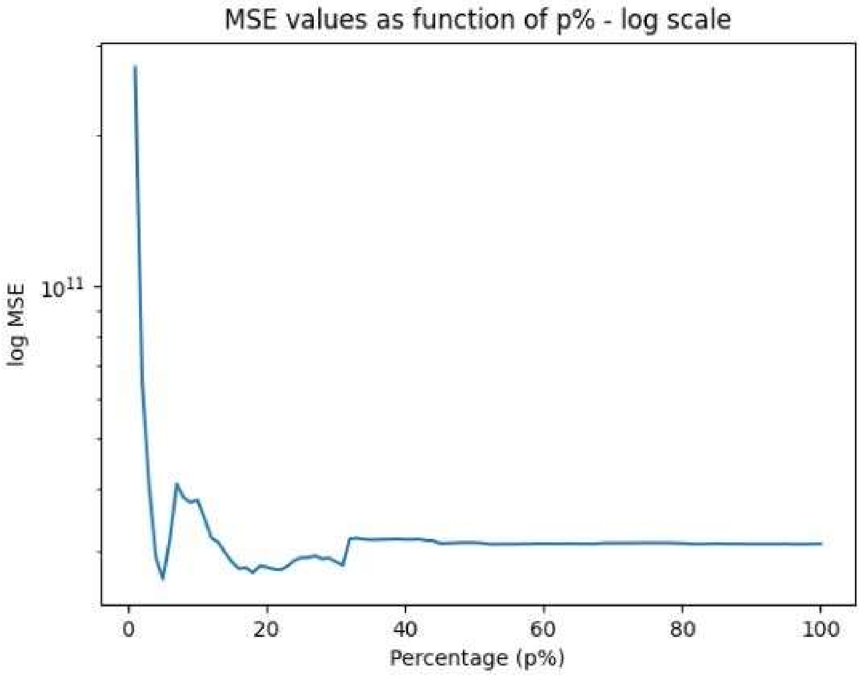
$$||\hat{w}||^2 = \sum_{i=1}^{d} \hat{w}_i^2 = \sum_{i=1}^{r} \hat{w}_i^2 + \sum_{i=r+1}^{d} \hat{w}_i^2 \leq \sum_{i=1}^{r} \overline{w}_i^2 + \sum_{i=r+1}^{d} \overline{w}_i^2 = ||\overline{w}||^2$$

Practical Questions

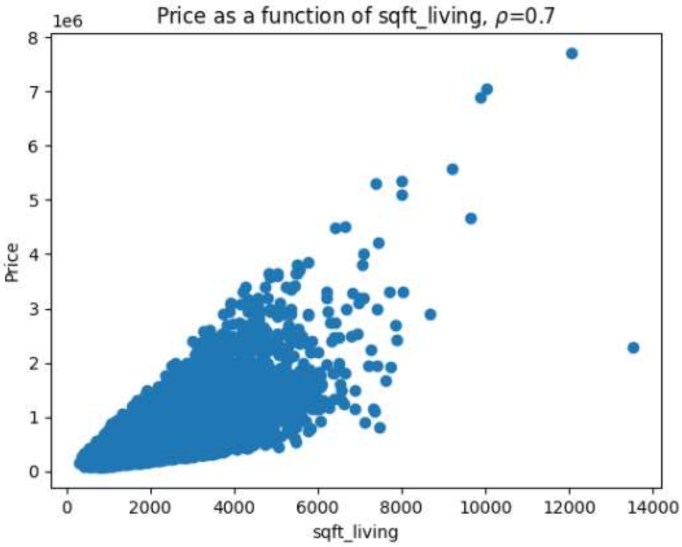
9. Code
10. Code
11. Code
12. Code
13. I choose the following features as categorical: lat, long, id and zip code, though in the preprocessing I have removed all of the features except the zip code, because they appeared redundant or with high correlation with zip code (therefore were not necessarily adding value to the fit)
14. A plot of the singular values in descending order:



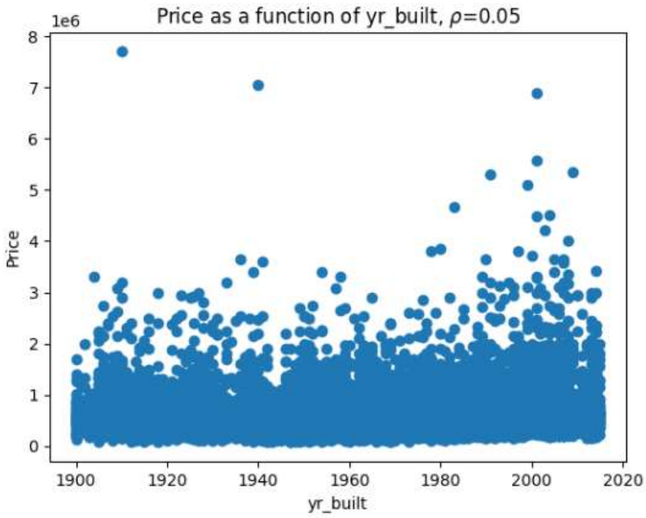
15. We can understand the importance of a feature by taking a look on its corresponding singular value. This is due to the fact that the singular values are correlated to the eigenvalues, and we know that large eigenvalue implies a more significant eigenvector, i.e., a more significant feature. If so, by observing the grape of Q.15, we can see that the values of the singular values drop fast, and from here it is understood that only a small portion of the features represents the price. In other words, while there are many features that were taken into consideration, only a few of them do correlate significantly with the price, therefore characterized with large singular values.
    Furthermore, it is not close to be singular because the singular values are larger than zero.
16. In the following graph we see the MSE value as a function of p%. some notes to take into consideration are:
    - The overall trend is that for larger data set the error is lower. This is correct up to some p%, for which is seems that the error is reaching some fixed value. This may be due to the fact that more information (i.e., more samples) does not necessarily indicates better fit – a plausible case is that we can not produce more information from an ever-growing data set.
    - An undesirable behavior occurs for the range $p \in \sim[5,25]$. Furthermore, when running the code multiple times, the behavior in this region is significantly different. This may be the result of small data, in which a fit is generated based on insufficient information.



17. First example – price as a function of square foot living (highly correlated)



Second example = price as a function of year built (highly uncorrelated)



On the one hand, larger living room corresponds to more expensive home, due to that the correlation between these two is high ($\rho \cong 0.7$). on the other hand, there is no apparent relation between the year built to the price, for the data portrayed in the second graph seems arbitrarily scattered. This is the reason the correlation is low ($\rho \cong 0.05$)