2022

Lecture 6: Linear Regression

Week 6

Lecturer: Tianyu Wang

1 Least Square Regression

Consider dataset $\{(x_i, y_i)\}_{i=1}^n$ ($x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$). A least square regression model, or a least square regressor, learns a model $f(x) = \theta^\top x$ by minimizing the following objective

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{2n} \sum_{i=1}^n \left(\theta^\top x_i - y_i\right)^2 .$$

Let

$$\ell(\theta) = \frac{1}{2} \sum_{i=1}^n \left(\theta^\top x_i - y_i\right)^2 = \frac{1}{2} \|\mathbf{X}\theta - \mathbf{y}\|_2^2,$$

where $\mathbf{X} = [x_1, x_2, \cdots, x_n]^\top$ and $\mathbf{y} = [y_1, y_2, \cdots, y_n]^\top$. This objective is convex in θ.

We take the gradient with respect to θ to get

$$\nabla \ell(\theta) = \mathbf{X}^\top (\mathbf{X}\theta - \mathbf{y}).$$

Setting $\nabla \ell(\theta) = 0$ gives $\mathbf{X}^\top \mathbf{X}\theta = \mathbf{X}^\top \mathbf{y}$, or $\theta = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$, provided that $\mathbf{X}^\top \mathbf{X}$ is invertible.

1.1 Ridge Regression

The loss for ridge regression is

$$\ell(\theta) = \frac{1}{2} \|\mathbf{X}\theta - \mathbf{y}\|_2^2 + \frac{\alpha}{2} \|\theta\|_2^2,$$

where α is a hyperparameter. The $\frac{\alpha}{2}\|\theta\|_2^2$ term is called the L_2-regularization term or the L_2-penalty term. The closed-form solution to ridge regression is

$$\theta = (\mathbf{X}^\top \mathbf{X} + \alpha I_d)^{-1} \mathbf{X}^\top \mathbf{y},$$

where I_d is the identity matrix of size $d \times d$.

1.1.1 Singular Value Decomposition (SVD)

Definition 1.1 (Singular Values)**.** For any matrix $A \in \mathbb{R}^{n\times m}$ and $i = 1, 2, \cdots, \min\{n, m\}$, let $\lambda_i(A^\top A)$ be the i-th eigenvalues of $A^\top A$. The i-th singular values of A is $\sigma_i(A) = \sqrt{\lambda_i(A^\top A)}$.

Definition 1.2 (Orthogonal Matrix)**.** A non-singular matrix $Q \in \mathbb{R}^{n\times n}$ is orthogonal if $Q^{-1} = Q^\top$.

Definition 1.3 (Singular Value Decomposition)**.** For any (non-zero) matrix $A \in \mathbb{R}^{n\times m}$, there exists a tuple of matrices (U, Σ, V) satisfying

- $A = U\Sigma V^\top$;
- (1) $U \in \mathbb{R}^{n\times n}$ and U is an orthogonal matrix, (2) $V \in \mathbb{R}^{m\times m}$ is an orthogonal matrix, (3) $\Sigma \in \mathbb{R}^{m\times n}$ is a matrix whose (i, i)-th entry is the i-th singular value of A, and all other entries are zero.

The decomposition $A = U\Sigma V^\top$ is called the singular value decomposition of A. The columns in U are called the left singular vectors of A, and the columns in V are called the right singular vectors of A.

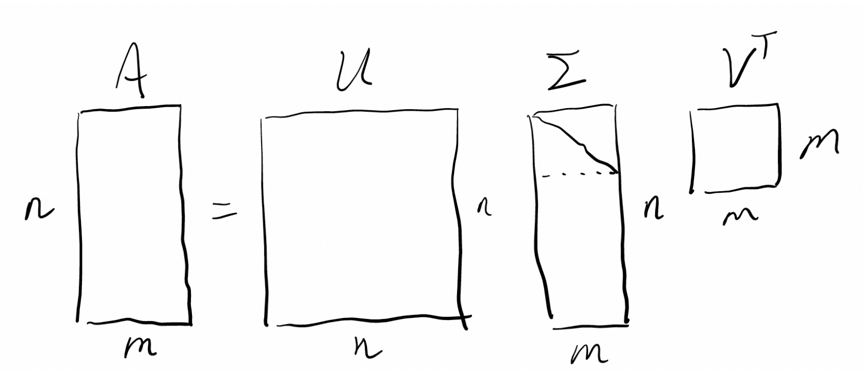

Figure 1: SVD illustration.

Proposition 1.4 (Some properties of SVD)**.** *Consider a matrix $A \in \mathbb{R}^{n\times m}$, and use the notations described in the caption of Figure 2. The SVD and compact SVD of A has the following properties.*

- *The property described in the caption of Figure 2.*
- *The rank of matrix A equals the number of non-zero singular values of the SVD.*
- *The columns of U are the eigenvectors of $AA^\top$.*
- *The columns of V are the eigenvectors of $A^\top A$.*

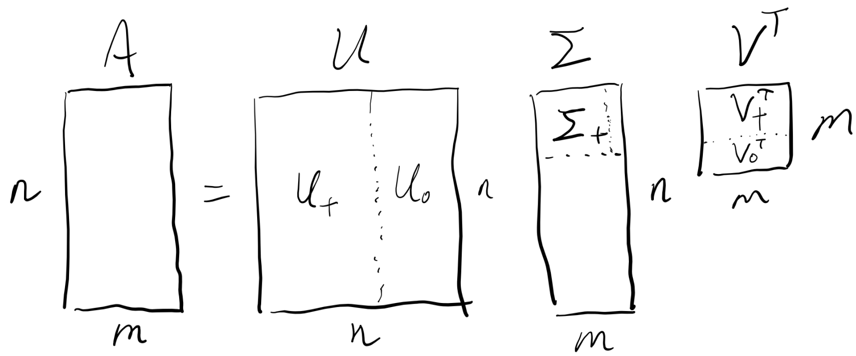

Figure 2: SVD illustration. Let $\Sigma_+ \in \mathbb{R}^{r\times r}$ be the square diagonal matrix of strictly positive singular values on its diagonal. The SVD gives the four fundamental spaces associated with a matrix. The columns of $U_+ \in \mathbb{R}^{n\times r}$ spans the column space of A; the columns of $U_0 \in \mathbb{R}^{n\times(n-r)}$ spans the left null space of A; the columns of $V_+ \in \mathbb{R}^{m\times r}$ (rows of $V_+^\top$) spans the row space of A; the columns of $V_0 \in \mathbb{R}^{m\times(m-r)}$ (rows of $V_0^\top$) spans the kernel space of A. It is easy to verify that $A = U\Sigma V^\top = U_+\Sigma_+V_+^\top$. We call $A = U_+\Sigma_+V_+^\top$ the compact SVD of A.

1.1.2 "Shrinkage" Effect via L_2-regularization

Let $\mathbf{X} = U\Sigma V^\top$ be the singular value decomposition of the data matrix $\mathbf{X}$. Let θ^{LS} be the solution of the least square regression objective, and let θ^R be the solution of the ridge regression objective. Let's take a closer look at θ^{LS} and θ^R. We have

$$\begin{aligned}
\theta^R &= \left(\mathbf{X}^\top\mathbf{X} + \alpha I_d\right)^{-1}\mathbf{X}^\top\mathbf{y} \\
&= \left(\left(U\Sigma V^\top\right)^\top\left(U\Sigma V^\top\right) + \alpha I_d\right)^{-1}\mathbf{X}^\top\mathbf{y} \\
&= \left(V\Sigma^\top\Sigma V^\top + \alpha I_d\right)^{-1}\mathbf{X}^\top\mathbf{y} && \text{(since } U^\top U = UU^\top = I_n) \\
&= \left(V\Sigma^\top\Sigma V^\top + \alpha VV^\top\right)^{-1}\mathbf{X}^\top\mathbf{y} && \text{(since } VV^\top = V^\top V = I_n) \\
&= \left(V\left(\Sigma^\top\Sigma + \alpha I_d\right)V^\top\right)^{-1}\mathbf{X}^\top\mathbf{y} \\
&= V\left(\Sigma^\top\Sigma + \alpha I_d\right)^{-1}V^\top\mathbf{X}^\top\mathbf{y}. && \text{(since } \left(QAQ^\top\right)^{-1} = QA^{-1}Q^\top \text{ as long as } Q \text{ is orthogonal)}
\end{aligned}$$

Note that $\Sigma^\top\Sigma$ is a square diagonal matrix. Let $\Sigma^\top\Sigma = diag(\lambda_1, \lambda_2, \cdots, \lambda_d)$. We have

$$\theta^R = diag\left(\frac{1}{\lambda_1+\alpha}, \frac{1}{\lambda_2+\alpha}, \cdots, \frac{1}{\lambda_d+\alpha}\right)\mathbf{X}^\top\mathbf{y}$$

Similarly, we have

$$\theta^{LS} = diag\left(\frac{1}{\lambda_1}, \frac{1}{\lambda_2}, \cdots, \frac{1}{\lambda_d}\right)\mathbf{X}^\top\mathbf{y}.$$

This means θ^R "shrinks" the values of θ^{LS}. This implies that the model given by θ^R is more conservative, and usually less likely to overfit.

1.2 Lasso Regression

The loss for lasso regression is

$$\ell(\theta) = \|\mathbf{X}\theta - \mathbf{y}\|_2^2 + \alpha\|\theta\|_1,$$

where α is a hyperparameter. The $\alpha\|\theta\|_1$ term is called the L_1-regularization term or the L_1-penalty term. Lasso regression can produce a sparse model, in which many entries of θ^{lasso} (the minimizer of the lasso regression loss) is zero.

Interpreting the objective as the Lagrangian of a contrained convex program, we get

$$\min_\theta \ \|\mathbf{X}\theta - \mathbf{y}\|_2^2 \qquad \text{subject to} \ \ \|\theta\|_1 \leq \alpha',$$

for some α'. Very likely, the optimal solution to this program lands at one of the vertices of $\{\theta : \|\theta\|_1 \leq \alpha'\}$, in which cases some entries of θ are zero.

2 Matrix Norms

We will discuss some matrix norms related to singular values. Consider $A \in \mathbb{R}^{n\times m}$. The Frobenius norm of A is

$$\begin{aligned} \|A\|_F &= \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}^2} \\ &= \sqrt{trace(A^\top A)} \\ &= \sqrt{\sum_{i=1}^{\min\{m,n\}} \lambda_i(A^\top A)} \\ &= \sqrt{\sum_{i=1}^{\min\{m,n\}} (\sigma_i(A))^2}. \end{aligned}$$

The Schatten p-norm (or p-Schatten norm) of a matrix A is

$$\|A\|_p = \left(\sum_{i=1}^{\min\{m,n\}} (\sigma_i(A))^p\right)^{1/p}.$$

The induced p-norm a matrix A (also written $\|A\|_p$) is

$$\|A\|_p = \sup_{x:x\neq 0} \frac{\|Ax\|_p}{\|x\|_p}.$$

Note. The Schatten p-norm and the induced p-norm are different. For example, the induced 2-norm equals the Schatten ∞-norm.

3 Some Applications of Singular Value Decomposition

Let $A = U\Sigma V^\top$ ($A \in \mathbb{R}^{n\times m}$) be the singular value decomposition of A. Let u_i be the columns of U and let v_i be the columns of V. We can write A as $A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$, where $r \le \min\{m, n\}$, and σ_i are the non-zero singular values of A.

3.1 Image Compression

Consider approximating a matrix $A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$ by $\widehat{A} = \sum_{i=1}^{s} \sigma_i u_i v_i^\top$ for some $s < r$. Let's look at the difference between A and $\widehat{A}$:

$$\|A - \widehat{A}\|_F = \|\sum_{i=s+1}^{r} \sigma_i u_i v_i^\top\|_F = \sqrt{\sum_{i=s+1}^{r} \sigma_i^2}.$$

One way for image compression is to do an SVD and take only several top singular values and singular vectors. If we keep top s singular values and top s singular vectors, we only need $O((n+m)s)$ spaces to store the image. In the homework, you will need to use a Python package to write a program for image compression via SVD.

There are many ways for image compression, and SVD is only one of them.

3.2 Pseudo-inverse and Least Square Linear Regression

Consider a matrix $A \in \mathbb{R}^{n\times m}$, and let $A = U_+\Sigma_+V_+^\top$ be its compact SVD. The pseudoinverse of A is $A^\dagger = V_+\Sigma_+^{-1}U_+^\top$. Note that pseudoinverse always exists, no matter whether the matrix is invertible or not.

As we have discussed before, the closed-form solution to least-square regression is

$$\theta = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y},$$

when $\mathbf{X}^\top\mathbf{X}$ is invertible. When $\mathbf{X}^\top\mathbf{X}$ is not invertible, we have

$$\theta = (\mathbf{X}^\top\mathbf{X})^{\dagger}\mathbf{X}^\top\mathbf{y}.$$

If the matrix $\mathbf{X}^\top\mathbf{X}$ is the not invertible, the prediction of the least square regression model is

$$\widehat{\mathbf{y}} = \mathbf{X}\theta = (\mathbf{X}^\top\mathbf{X})^{\dagger}\mathbf{X}^\top\mathbf{y}.$$

Let $\mathbf{X} = U_+\Sigma_+V_+^\top$ ($\Sigma_+ \in \mathbb{R}^{r\times r}$) be the compact SVD of $\mathbf{X}$. We have

$$\begin{aligned}
\widehat{\mathbf{y}} &= U_+\Sigma_+V_+^\top \left(V_+\Sigma_+^{-1}U_+^\top U_+\Sigma_+^{-1}V_+^\top\right) V_+\Sigma_+U_+^\top\mathbf{y} \\
&= U_+U_+^\top\mathbf{y}. \qquad\qquad \text{(Note that } U_+^\top U_+ = V_+^\top V_+ = I_r.\text{)}
\end{aligned}$$

The columns of U_+ are orthonormal vectors. Let $u_{+,i}$ be the i-th column of U_+. Thus we have

$$\widehat{\mathbf{y}} = U_+U_+^\top\mathbf{y} = \sum_{i=1}^{r} u_{+,i}u_{+,i}^\top\mathbf{y},$$

which is the projection of $\mathbf{y}$ onto the column space of U_+.

3.3 Principal Component Analysis (PCA)

Consider a data matrix $\mathbf{X}$, and let $\mathbf{w}_{(0)} = 0$. The k-th principle component of $\mathbf{X}$ is

$$\mathbf{w}_{(k)} = \arg \max_{\mathbf{w}:\|w\|_2=1} \|\mathbf{X}_k \mathbf{w}\|_2^2,$$

where

$$\mathbf{X}_k = \mathbf{X} - \sum_{i=0}^{k-1} \mathbf{X}\mathbf{w}_{(i)}\mathbf{w}_{(i)}^\top.$$

By this objective, $\mathbf{w}_{(1)}$ is the direction along which the data points in $\mathbf{X}$ vary the most. Similarly, $\mathbf{w}_{(k)}$ is the direction along which the data points in $\mathbf{X}_k$ vary the most. Note that $\mathbf{w}_{(i)}\mathbf{w}_{(i)}^\top$ is the projection matrix as discussed previously. Thus $\mathbf{X}_k$ has removed the components from $\mathbf{w}_{(1)}, \mathbf{w}_{(2)}, \cdots, \mathbf{w}_{(k-1)}$.

Note that

$$\begin{aligned} \mathbf{w}_{(1)} &= \arg \max_{\mathbf{w}:\|w\|_2=1} \|\mathbf{X}\mathbf{w}\|_2^2 \\ &= \arg \max_{\mathbf{w}:\|w\|_2=1} \mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w}. \end{aligned}$$

This means $\mathbf{w}_{(1)}$ is the top eigenvalue of $\mathbf{X}^\top\mathbf{X}$ (homework exercise), which is the top right singular vector of $\mathbf{X}$. Similarly, $\mathbf{w}_{(k)}$ is the top eigenvalue of $\mathbf{X}_k^\top \mathbf{X}_k$.

An illustration of SVD and PCA is given in Figure 3.

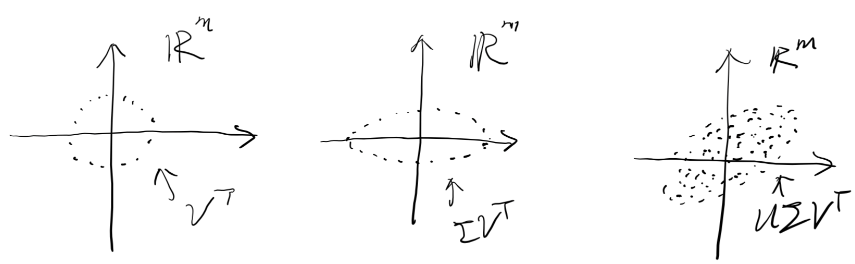

Figure 3: Illustration of PCA. We can recover the data matrix $\mathbf{X} = U\Sigma V^\top$ ($\mathbf{X} \in \mathbb{R}^{n\times n}$) from left to right. The rightmost subfigure plots the data points in $\mathbb{R}^m$.

Acknowledgement

Reference: Machine Learning: A Probabilistic Perspective by Kevin Murphy. A thank you to wikipedia contributors.