Lecture 5 Linear regression II and K-Means Clustering

Recall that in a linear regression problem, we aim to identify a regression model

$$\hat{f}(\vec{x}_i; \vec{\beta}) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im}$$

for any given vector $\vec{x}_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{im} \end{bmatrix} \in \mathbb{R}^m$,

such that with n paired observations

$$(\vec{x}_1, y_1), \quad (\vec{x}_2, y_2), \quad \cdots, \quad (\vec{x}_n, y_n),$$

the following loss function is minimized,

$$J(\vec{\beta}) = \sum_{i=1}^{n} \big(\underbrace{y_i - \beta_0 - \sum_{j=1}^{m} x_{ij}\beta_j}_{\text{residual}} \big)^2 = \| \vec{y} - \underline{X}\vec{\beta} \|_2^2$$

where

$$\underline{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix}_{n \times (m+1)} \quad \text{(design matrix)}$$

$\vec{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_m \end{bmatrix} \in \mathbb{R}^{m+1}$ is the parameter vector.

$\vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n$. response variable

Last time, we showed that the unique minimizer of the least squares problem

$$\min_{\vec{\beta}\in\mathbb{R}^{m+1}} \|\vec{y} - \underline{\bar{X}}\vec{\beta}\|_2^2 = \min_{\vec{\beta}\in\mathbb{R}^{m+1}} J(\vec{\beta})$$

is given by $\underbrace{\underline{\bar{X}}^T}_{(m+1)\times n} \underbrace{\underline{\bar{X}}}_{n\times(m+1)} \vec{\beta}^* = \underline{\bar{X}}^T \vec{y} \Rightarrow \vec{\beta}^* = (\underline{\bar{X}}^T\underline{\bar{X}})^{-1}\underline{\bar{X}}^T\vec{y}$

where $n \geq m+1$, $\underline{\bar{X}}$ is full column rank.

Here is an alternative derivation:

Let $r_i = y_i - \sum_{j=0}^{m} \underline{\bar{X}}_{ij}\beta_j$ ①, $i=1,\cdots,n$

then $J(\vec{\beta}) = \sum_{i=1}^{n} r_i^2$

we want $\left.\partial J/\partial\beta_k\right|_{\vec{\beta}=\vec{\beta}^*} = 0$ for $k=0,1,\cdots,m$

That is, $\sum_{i=1}^{n} 2r_i \left.\frac{\partial r_i}{\partial\beta_k}\right|_{\vec{\beta}=\vec{\beta}^*} = 0$, ② for $k=0,1,\cdots,m$

From ①, we have $\left.\frac{\partial r_i}{\partial\beta_k}\right|_{\vec{\beta}=\vec{\beta}^*} = -\underline{\bar{X}}_{ik}$ ③

Substitute ① and ③ into ②: $\sum_{i=1}^{n}[y_i - \sum_{j=0}^{m}\underline{\bar{X}}_{ij}\beta_j^*](-\underline{\bar{X}}_{ik}) = 0$

$k=0,\cdots,m$

That is, $\sum_{i=1}^{n}\sum_{j=0}^{m}\underline{\bar{X}}_{ij}\underline{\bar{X}}_{ik}\beta_j^* = \sum_{i=1}^{n} y_i\underline{\bar{X}}_{ik}$ $k=0,\cdots,m$

$$\Rightarrow \sum_{i=1}^{n}\sum_{j=0}^{m}\underline{\bar{X}}^T_{ji}\underline{\bar{X}}_{ik}\beta_j^* = \sum_{i=1}^{n}\underline{\bar{X}}^T_{ki}y_i$$

This is equivalent to $\underline{\bar{X}}^T\underline{\bar{X}}\vec{\beta} = \underline{\bar{X}}^T\vec{y}$ (normal equation)

§ K-means clustering

References: ① James et. al. An intro to statistical learning
Sec 12.5.3
② Calvetti & Somersalo. Mathematics of data science: A Computational approach to clustering and classification. SIAM 2021.

Clustering is a common approach in unsupervised learning, where the goal is to organize a collection of data points $\vec{x}_i \in \mathbb{R}^m$ $(i=1,\cdots,n)$ into k groups.

so that the vectors within each group (called a cluster) are similar to each other based on a defined measure of distance in $\mathbb{R}^m$.

See Jupyter notebook for an example.

Definition: Given a set of n data vectors
$$D = \{\vec{x}_i \in \mathbb{R}^m \mid i=1,\cdots,n\}$$
we arrange these vectors into k distinct clusters,
$$D_l = \{\vec{x}_j \mid j \in I_l\}, \quad l=1,2,\cdots,k.$$
where the index sets I_l satisfy
$$\bigcup_{l=1}^{k} I_l = \{1,\cdots,n\}, \quad \text{and}$$
$$I_l \cap I_j = \phi, \quad l \neq j \quad \text{where } l,j = 1,\cdots,k.$$