ESE 402/542 Recitation 7: Linear Regression

Outline

1. Linear regression in matrix form
2. Polynomial fitting
3. Linear regression and maximum likelihood

Linear Regression in Matrix Form

1. Given (x_i, y_i) pairs, want to fit a line that predicts y from x optimally (least-squares sense)
2. In scalar case, $x_i \in \mathbb{R}$, and assume that the $y_i \in \mathbb{R}$ are generated via $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{dist}(0, \sigma^2)$
3. In vector case, $\boldsymbol{x}_i \in \mathbb{R}^d$. Assume that the $y_i \in \mathbb{R}$ are generated via $y_i = \beta_0 + \beta_1 \boldsymbol{x}_{i,1} + \cdots + \beta_d \boldsymbol{x}_{i,d} + \epsilon_i$
4. Stacking all n samples into a matrix,

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 \boldsymbol{x}_{1,1} + \cdots + \beta_d \boldsymbol{x}_{1,d} \\ \vdots \\ \beta_0 + \beta_1 \boldsymbol{x}_{n,1} + \cdots + \beta_d \boldsymbol{x}_{n,d} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{1}$$

$$= \begin{bmatrix} 1 & \boldsymbol{x}_{1,1} & \dots & \boldsymbol{x}_{1,d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \boldsymbol{x}_{n,1} & \dots & \boldsymbol{x}_{n,d} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_d \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{2}$$

Linear Regression in Matrix Form

In matrix form, we can assume our data has been generated

$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \epsilon$, where $\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$, $\boldsymbol{X} = \begin{bmatrix} 1 & \boldsymbol{x}_{1,1} & \dots & \boldsymbol{x}_{1,d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \boldsymbol{x}_{n,1} & \dots & \boldsymbol{x}_{n,d} \end{bmatrix}$,

$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_d \end{bmatrix}$, and $\epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$. Note that $\mathbb{E}[\epsilon] = \vec{0}$ and $\mathbf{cov}(\epsilon) = \sigma^2 I_n$.

We want to minimize RSS and solve $\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\beta\|^2$.
Taking derivative w.r.t. $\boldsymbol{\beta}$ and setting to 0, we get

$$\hat{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

Linear Regression and Maximum Likelihood

Natural question: when is the above linear estimator provably "good"?

1. Above, we assumed $\epsilon \sim \mathrm{dist}(0, \sigma^2 \boldsymbol{I}_n)$
2. Additionally assume that $\epsilon \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$. Then, $y_i | \boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\beta}^\top \boldsymbol{x}_i, \sigma^2)$. Can re-derive least-squares linear regression through MLE
3. Solve $\hat{\boldsymbol{\beta}} = \operatorname{argmax}_{\boldsymbol{\beta}} \Pr(\boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_n, y_n | \boldsymbol{\beta}) = \operatorname{argmax}_{\boldsymbol{\beta}} \prod_{i=1}^n \Pr(y_i | x_i, \boldsymbol{\beta}) = \operatorname{argmax}_{\boldsymbol{\beta}} \sum_{i=1}^n \log \Pr(y_i | x_i, \boldsymbol{\beta}) = \operatorname{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^n (\boldsymbol{\beta}^\top x_i - y_i)^2 = \operatorname{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|^2$ which is the same formulation as we had above

Polynomial Fitting

Special case of linear regression: given the degree d of the polynomial, then we want to fit a model of the form

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_d x^d.$$

Note that this is a special case of linear regression by setting $\boldsymbol{x}_k = x^k$ for $k = 0, \ldots, d$. Given n samples of the form

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_d x_i^d + \epsilon_i,$$

we turn this into matrix form

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & \cdots & x_1^d \\ 1 & x_2 & \cdots & x_2^d \\ \vdots & & & \\ 1 & x_n & \cdots & x_n^d \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Polynomial Fitting

We can now solve for $\hat{\beta}_0, \dots, \hat{\beta}_d$ like earlier. The matrix

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & \cdots & x_1^d \\ 1 & x_2 & \cdots & x_2^d \\ \vdots \\ 1 & x_n & \cdots & x_n^d \end{bmatrix},$$

is also known as the Vandermonde matrix.