CSC 412 Machine Learning and Knowledge Discovery

Linear Regression

1 Notations

- In machine learning, vectors are often represented as column vectors
- Lowercase italic font is for scalar values, such as m or $y^{(i)}$
- Lowercase bold font is for vectors, such as $\boldsymbol{x}^{(i)}$
- Uppercase bold font is for matrices, such as $\boldsymbol{X}$
- m is the number of instances (rows; samples)
- n is the number of features (columns)
- $\boldsymbol{x}^{(i)}$ is a vector of all the feature values (excluding the label) of the i^{th} instance (the i^{th} row)
- $\boldsymbol{x}_j$ is the j^{th} feature value (the j^{th} column)
- $\boldsymbol{X}$ is a matrix containing all the feature values (excluding labels) of all instances in the dataset. There is one row per instance, and the i^{th} row is equal to the transpose of $\boldsymbol{x}^{(i)}$, noted $(\boldsymbol{x}^{(i)})^\intercal$

$$\boldsymbol{X} = \begin{bmatrix} \text{———} & (\boldsymbol{x}^{(1)})^\intercal & \text{———} \\ \text{———} & (\boldsymbol{x}^{(2)})^\intercal & \text{———} \\ \text{———} & (\boldsymbol{x}^{(3)})^\intercal & \text{———} \\ & \vdots & \\ \text{———} & (\boldsymbol{x}^{(m)})^\intercal & \text{———} \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & \cdots & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & \cdots & x_n^{(2)} \\ x_1^{(3)} & x_2^{(3)} & x_3^{(3)} & \cdots & x_n^{(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1^{(m)} & x_2^{(m)} & x_3^{(m)} & \cdots & x_n^{(m)} \end{bmatrix}$$

- $\boldsymbol{y}$ is the label vector (all the desired output values; ground truth)
- $y^{(i)}$ is $\boldsymbol{x}^{(i)}$'s label (the desired output value for that instance)

$$\boldsymbol{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ y^{(3)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

- $\hat{\boldsymbol{y}}$ is the predicted vector

2 Model Prediction

2.1 One Instance

$$\hat{y} = \sum_{j=0}^{n} \theta_j \cdot x_j = \theta_0 \cdot x_0 + \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \cdots + \theta_n \cdot x_n$$

with $x_0 = 1$

- θ_j is the j^{th} model parameter, including the bias term θ_0 and the feature weights θ_1, θ_2, ... , θ_n

2.2 Vectorized Form

$$\boxed{\hat{y} = h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^\intercal \boldsymbol{x}} = \begin{bmatrix} \theta_0 & \theta_1 & \theta_2 & \cdots & \theta_n \end{bmatrix} \cdot \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

- $h_{\boldsymbol{\theta}}$ is the hypothesis function, using the model parameters $\boldsymbol{\theta}$

- $\boldsymbol{\theta}$ is the model's parameter vector, containing the bias term θ_0 and the feature weights θ_1 to θ_n

- $\boldsymbol{x}$ is the instance's feature vector, containing x_0 to x_n, with x_0 always equal to 1.

2.3 Matrixed Form

$$\boxed{\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\theta}} = \begin{bmatrix} \text{———} & (\boldsymbol{x}^{(1)})^\intercal & \text{———} \\ \text{———} & (\boldsymbol{x}^{(2)})^\intercal & \text{———} \\ \text{———} & (\boldsymbol{x}^{(3)})^\intercal & \text{———} \\ & \vdots & \\ \text{———} & (\boldsymbol{x}^{(m)})^\intercal & \text{———} \end{bmatrix} \cdot \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_n^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_n^{(2)} \\ 1 & x_1^{(3)} & x_2^{(3)} & \cdots & x_n^{(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(m)} & x_2^{(m)} & \cdots & x_n^{(m)} \end{bmatrix} \cdot \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix} = \begin{bmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \hat{y}^{(3)} \\ \vdots \\ \hat{y}^{(m)} \end{bmatrix}$$

3 Cost Function

3.1 Mean Squared Error (MSE)

$$\begin{aligned} J(\boldsymbol{\theta}) &= \frac{1}{m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)} \right)^2 \\ &= \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{\theta}^\intercal \boldsymbol{x}^{(i)} - y^{(i)} \right)^2 \\ &= \frac{1}{m} \left(\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y} \right)^\intercal \left(\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y} \right) \end{aligned}$$

3.2 Gradient Vector

$$
\begin{aligned}
\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= \nabla_{\boldsymbol{\theta}} \Big(\frac{1}{m} \cdot (\boldsymbol{X\theta} - \boldsymbol{y})^{\intercal} \cdot (\boldsymbol{X\theta} - \boldsymbol{y}) \Big) \\
&= \frac{1}{m} \cdot \nabla_{\boldsymbol{\theta}} \Big((\boldsymbol{X\theta} - \boldsymbol{y})^{\intercal} \cdot (\boldsymbol{X\theta} - \boldsymbol{y}) \Big) \\
&= \frac{1}{m} \cdot \nabla_{\boldsymbol{\theta}} \Big((\boldsymbol{\theta}^{\intercal}\boldsymbol{X}^{\intercal} - \boldsymbol{y}^{\intercal}) \cdot (\boldsymbol{X\theta} - \boldsymbol{y}) \Big) \quad \text{(1) (2)} \\
&= \frac{1}{m} \cdot \nabla_{\boldsymbol{\theta}} \Big(\boldsymbol{\theta}^{\intercal}\boldsymbol{X}^{\intercal}\boldsymbol{X\theta} - \boldsymbol{\theta}^{\intercal}\boldsymbol{X}^{\intercal}\boldsymbol{y} - \boldsymbol{y}^{\intercal}\boldsymbol{X\theta} + \boldsymbol{y}^{\intercal}\boldsymbol{y} \Big) \\
&= \frac{1}{m} \cdot \Big(\nabla_{\boldsymbol{\theta}} (\boldsymbol{\theta}^{\intercal}\boldsymbol{X}^{\intercal}\boldsymbol{X\theta}) - \nabla_{\boldsymbol{\theta}} (\boldsymbol{\theta}^{\intercal}\boldsymbol{X}^{\intercal}\boldsymbol{y}) - \nabla_{\boldsymbol{\theta}} (\boldsymbol{y}^{\intercal}\boldsymbol{X\theta}) + \nabla_{\boldsymbol{\theta}} (\boldsymbol{y}^{\intercal}\boldsymbol{y}) \Big) \\
&= \frac{1}{m} \cdot \Big(2\boldsymbol{X}^{\intercal}\boldsymbol{X\theta} - \boldsymbol{X}^{\intercal}\boldsymbol{y} - \boldsymbol{X}^{\intercal}\boldsymbol{y} \Big) \quad \text{(5) (3) (4)} \\
&= \frac{2}{m} \cdot \Big(\boldsymbol{X}^{\intercal}\boldsymbol{X\theta} - \boldsymbol{X}^{\intercal}\boldsymbol{y} \Big) \\
&= \frac{2}{m} \cdot \boldsymbol{X}^{\intercal} \cdot (\boldsymbol{X\theta} - \boldsymbol{y})
\end{aligned}
$$

So,

$$\boxed{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{2}{m} \boldsymbol{X}^{\intercal} (\boldsymbol{X\theta} - \boldsymbol{y})}$$

Matrix Formulae:
(1) $(\boldsymbol{A} + \boldsymbol{B})^{\intercal} = \boldsymbol{A}^{\intercal} + \boldsymbol{B}^{\intercal}$
(2) $(\boldsymbol{AB})^{\intercal} = \boldsymbol{B}^{\intercal}\boldsymbol{A}^{\intercal}$
(3) $\nabla_{\boldsymbol{x}} (\boldsymbol{x}^{\intercal}\boldsymbol{A}) = \boldsymbol{A}$
(4) $\nabla_{\boldsymbol{x}} (\boldsymbol{Ax}) = \boldsymbol{A}^{\intercal}$
(5) $\nabla_{\boldsymbol{x}} (\boldsymbol{x}^{\intercal}\boldsymbol{Ax}) = (\boldsymbol{A} + \boldsymbol{A}^{\intercal})\boldsymbol{x}$

3.3 Normal Equation (Closed-Form Solution)

Let

$$
\begin{aligned}
\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) &= 0 \\
\frac{2}{m} \cdot \Big(\boldsymbol{X}^{\intercal}\boldsymbol{X\theta} - \boldsymbol{X}^{\intercal}\boldsymbol{y} \Big) &= 0 \\
\boldsymbol{X}^{\intercal}\boldsymbol{X\theta} - \boldsymbol{X}^{\intercal}\boldsymbol{y} &= 0 \\
\boldsymbol{X}^{\intercal}\boldsymbol{X\theta} &= \boldsymbol{X}^{\intercal}\boldsymbol{y} \\
\boldsymbol{\theta} &= (\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1}\boldsymbol{X}^{\intercal}\boldsymbol{y}
\end{aligned}
$$

So,

$$\boxed{\hat{\boldsymbol{\theta}} = (\boldsymbol{X}^{\intercal}\boldsymbol{X})^{-1}\boldsymbol{X}^{\intercal}\boldsymbol{y}}$$

3.4 Gradient Descent

$$\boxed{\boldsymbol{\theta}^{(\text{next step})} := \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})}$$