*** Applied Machine Learning Fundamentals *** Regression

Daniel Wehner

SAP SE

August 12, 2019

Agenda August 12, 2019

Section:

Introduction to Regression

Regression

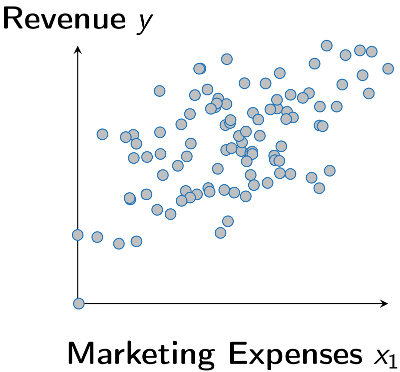
Type of target variable	Continuous
Type of training information	Supervised
Example Availability	Batch learning

Algorithm sketch: Given the training data $\mathcal{D}$ the algorithm derives a function of the type

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \theta_0 + \theta_1 x_1 + \cdots + \theta_{m+1} x_m \qquad \boldsymbol{x} \in \mathbb{R}^m, \boldsymbol{\theta} \in \mathbb{R}^{m+1} \tag{1}$$

from the data. $\boldsymbol{\theta}$ is the parameter vector containing the coefficients to be estimated by the regression algorithm. Once $\boldsymbol{\theta}$ is learned it can be used for prediction.

Example Data Set: Revenues



- Find a linear function:

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \theta_0 + \theta_1 x_1 + \cdots + \theta_{m+1} x_m$$

- Usually: $x_0 = 1$:

$$\widehat{\boldsymbol{x}} \in \mathbb{R}^{m+1} = [1 \ \boldsymbol{x}]^\intercal$$

$$h_{\boldsymbol{\theta}}(\widehat{\boldsymbol{x}}) = \sum_{j=0}^{m+1} \theta_j x_j = \boldsymbol{\theta}^\intercal \widehat{\boldsymbol{x}}$$

Error Function for Regression

- In order to know how good the function fits we need an error function $\mathcal{J}(\boldsymbol{\theta})$:

$$\mathcal{J}(\boldsymbol{\theta}) = \frac{1}{2n} \sum_{i=1}^{n} \left(h_{\boldsymbol{\theta}}(\widehat{\boldsymbol{x}}^{(i)}) - y^{(i)}\right)^2 \tag{2}$$

- We want to minimize $\mathcal{J}(\boldsymbol{\theta})$:

$$\min_{\boldsymbol{\theta}} \frac{1}{2n} \sum_{i=1}^{n} \left(h_{\boldsymbol{\theta}}(\widehat{\boldsymbol{x}}^{(i)}) - y^{(i)}\right)^2$$

- This is ordinary least squares (OLS)

Error Function Intuition

Closed-Form Solutions

- Usual approach (for two unknowns): Calculate θ_0 and θ_1 according to

sample mean $\overline{x}$

$$\theta_0 = \overline{y} - \theta_2 \overline{x} \qquad \theta_1 = \frac{\sum_{i=1}^{n} (x^{(i)} - \overline{x}) \cdot (y^{(i)} - \overline{y})}{\sum_{i=1}^{n} (x^{(i)} - \overline{x})^2} \tag{3}$$

- **'Normal equation'** (scales to arbitrary dimensions):

$$\boldsymbol{\theta} = \underbrace{(\widehat{\boldsymbol{X}}^\intercal \widehat{\boldsymbol{X}})^{-1} \widehat{\boldsymbol{X}}^\intercal}_{\text{Moore-Penrose pseudo-inverse}} \boldsymbol{y} \tag{4}$$

$\widehat{\boldsymbol{X}}$ is called 'design matrix' or 'regressor matrix'

Design Matrix / Regressor Matrix

- The design matrix $\hat{\boldsymbol{X}} \in \mathbb{R}^{n \times (m+1)}$ looks as follows:

$$\hat{\boldsymbol{X}} = \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_m^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_m^{(2)} \\ 1 & x_1^{(3)} & x_2^{(3)} & \cdots & x_m^{(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(n)} & x_2^{(n)} & \cdots & x_m^{(n)} \end{pmatrix} \tag{5}$$

In the following

$\hat{\boldsymbol{X}} \equiv \boldsymbol{X}$

- And the $n \times 1$ label vector:

$$\boldsymbol{y} = \left(y^{(1)}, y^{(2)}, y^{(3)}, \dots, y^{(n)}\right)^{\mathsf{T}}$$

Derivation of the Normal Equation

- The derivation involves a bit of linear algebra
- Step ❶: Rewrite $\mathcal{J}(\boldsymbol{\theta})$ in matrix-vector notation:

$$\begin{aligned}
\mathcal{J}(\boldsymbol{\theta}) &= \frac{1}{2}(\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y})^\intercal(\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}) \\
&= ((\boldsymbol{X}\boldsymbol{\theta})^\intercal - \boldsymbol{y}^\intercal)(\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{y}) \\
&= (\boldsymbol{X}\boldsymbol{\theta})^\intercal\boldsymbol{X}\boldsymbol{\theta} - (\boldsymbol{X}\boldsymbol{\theta})^\intercal\boldsymbol{y} - \boldsymbol{y}^\intercal(\boldsymbol{X}\boldsymbol{\theta}) + \boldsymbol{y}^\intercal\boldsymbol{y} \\
&= \boldsymbol{\theta}^\intercal\boldsymbol{X}^\intercal\boldsymbol{X}\boldsymbol{\theta} - 2(\boldsymbol{X}\boldsymbol{\theta})^\intercal\boldsymbol{y} + \boldsymbol{y}^\intercal\boldsymbol{y}
\end{aligned}$$

- To be continued...

Derivation of the Normal Equation (Ctd.)

- Step ❷: Calculate the derivative of $\mathcal{J}(\boldsymbol{\theta})$ and set it to zero:

$$\nabla_{\boldsymbol{\theta}} \mathcal{J}(\boldsymbol{\theta}) = 2\boldsymbol{X}^\intercal \boldsymbol{X}\boldsymbol{\theta} - 2\boldsymbol{X}^\intercal \boldsymbol{y} \overset{!}{=} \mathbf{0}$$
$$\Leftrightarrow \boldsymbol{X}^\intercal \boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{X}^\intercal \boldsymbol{y}$$

- If $\boldsymbol{X}^\intercal \boldsymbol{X}$ is invertible, we can multiply both sides by $(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}$:

Normal equation:

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

Summary

Lecture Overview

Unit I: Machine Learning Introduction

Self-Test Questions

Recommended Literature and further Reading

Thank you very much for the attention!

Topic: *** Applied Machine Learning Fundamentals *** Regression
Date: August 12, 2019

Contact:
Daniel Wehner (D062271)
SAP SE
daniel.wehner@sap.com

Do you have any questions?