Artificial Neural Networks
Lecture Notes

Prof. Dr. Sen Cheng

Winter Semester 2019/20

3 Regression

3.1 Univariate linear regression

Simplest case of regression.

Data: one independent variable $x \in R$, and one dependent variable $y \in R$.
Model class: linear function, i.e.,

$$f(x) = mx + b \tag{1}$$

The parameters of the model are $\theta = (m, b)$.
Loss function: l_2-norm, or summed squared errors,

$$\begin{aligned} L(\theta, X, Y) &= \sum_{i=1}^{N} (y_i - f(x_i))^2 && (2)\\ &= \sum_{i=1}^{N} (y_i - mx_i - b)^2 && (3) \end{aligned}$$

This is why this method is also known as the *minimum least square* (MLS) method.
Optimization: Linear regression is simple enough that we can derive an analytical solution, i.e., the parameters that will minimize the loss function.

The loss function is minimized for values of m and b such that $\frac{\partial}{\partial b} L = 0$ and $\frac{\partial}{\partial m} L = 0$. The first condition gives us

$$\sum_{i=1}^{N} -2(y_i - mx_i - b) = 0 \tag{4a}$$

$$\sum y_i - m \sum x_i - Nb = 0 \tag{4b}$$

which results in

$$b = \frac{1}{N} \sum y_i - m \frac{1}{N} \sum x_i \tag{5}$$

The second condition, $\frac{\partial}{\partial m} L = 0$, leads to

$$\sum_{i=1}^{N} -2x_i(y_i - mx_i - b) = 0 \tag{6}$$

substituting the value of b from Eq. 5 in here, we get

$$\sum x_i y_i - m \sum x_i^2 - \left(\sum x_i\right) \left(\frac{1}{N} \sum y_i - m \frac{1}{N} \sum x_i \right) = 0 \tag{7}$$

$$\sum x_i y_i - \frac{1}{N} \left(\sum x_i\right) \left(\sum y_i\right) = m \left[\sum x_i^2 - \frac{1}{N} \left(\sum x_i\right)^2 \right] \tag{8}$$

$$m = \frac{\sum x_i y_i - \frac{1}{N}(\sum x_i)(\sum y_i)}{\sum x_i^2 - \frac{1}{N}(\sum x_i)^2} \tag{9}$$

Equation 5 and 9 give us the parameter values that will minimize the loss. A more memorable version of these equations is

$$m = \frac{\text{cov}(X,Y)}{\sigma^2(X)} \tag{10}$$

$$b = \bar{y} - m\bar{x} \tag{11}$$

where the covariance $\text{cov}(X,Y) = \frac{1}{N}\sum x_i y_i - \frac{1}{N^2}(\sum x_i)(\sum y_i)$.

3.2 Goodness of fit

Linear regression is guaranteed to find the linear function that best accounts for the data, but that doesn't mean much, if a linear function is not a good description of the data. How would you know? There are many methods. The simplest one is to look at the data and the fit line.

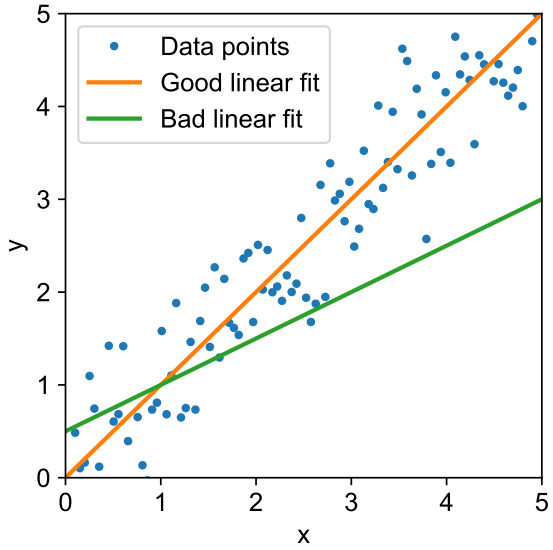

Figure 1: Good linear fit and bad linear fit

Explained variance R^2: A quantitative mesure of the quality of fit is the fraction of the variance in the data SS_{data} that can be accounted for by the model SS_{reg}.

$$SS_{data} = \sum_{i=1}^{N} (y_i - \bar{y})^2 \tag{12}$$

$$SS_{reg} = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2 \tag{13}$$

$$R^2 = \frac{SS_{reg}}{SS_{data}} \tag{14}$$

R^2 is called the explained variance, because the total variance in the data can be partitioned into two components, the variance predicted by the regression estimates, SS_{reg}, and the deviation between model and data $SS_{err} = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$

$$SS_{data} = SS_{reg} + SS_{err} \tag{15}$$

We can show that the sum of squares of the deviations can be partitioned as in Eq. 15 as follows :

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \tag{16a}$$

$$\underbrace{\sum_{i=1}^{N}(y_i-\bar{y})^2}_{SS_{data}} = \underbrace{\sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2}_{SS_{reg}} + \underbrace{\sum_{i=1}^{N}(y_i-\hat{y}_i)^2}_{SS_{err}} + 2\sum_{i=1}^{N}(\hat{y}_i-\bar{y})(y_i-\hat{y}_i) \tag{16b}$$

Inserting $\hat{y} = mx + b$ and splitting up the last term, we obtain

$$\sum_{i=1}^{N}(\hat{y}_i-\bar{y})(y_i-\hat{y}_i) = \sum_{i=1}^{N}(mx_i+b)(y_i-mx_i-b) - \sum_{i=1}^{N}\bar{y}(y_i-mx_i-b) \tag{16c}$$

$$= m\underbrace{\sum_{i=1}^{N}x_i(y_i-mx_i-b)}_{=0 \quad \text{(Eq. 6)}} + (b-\bar{y})\underbrace{\sum_{i=1}^{N}(y_i-mx_i-b)}_{=0 \quad \text{(Eq. 4a)}} \tag{16d}$$

Therefore

$$SS_{data} = SS_{reg} + SS_{err} \tag{16e}$$

Residual plot: calculate the residuals and plot them vs. the independent variable (Fig. 2).

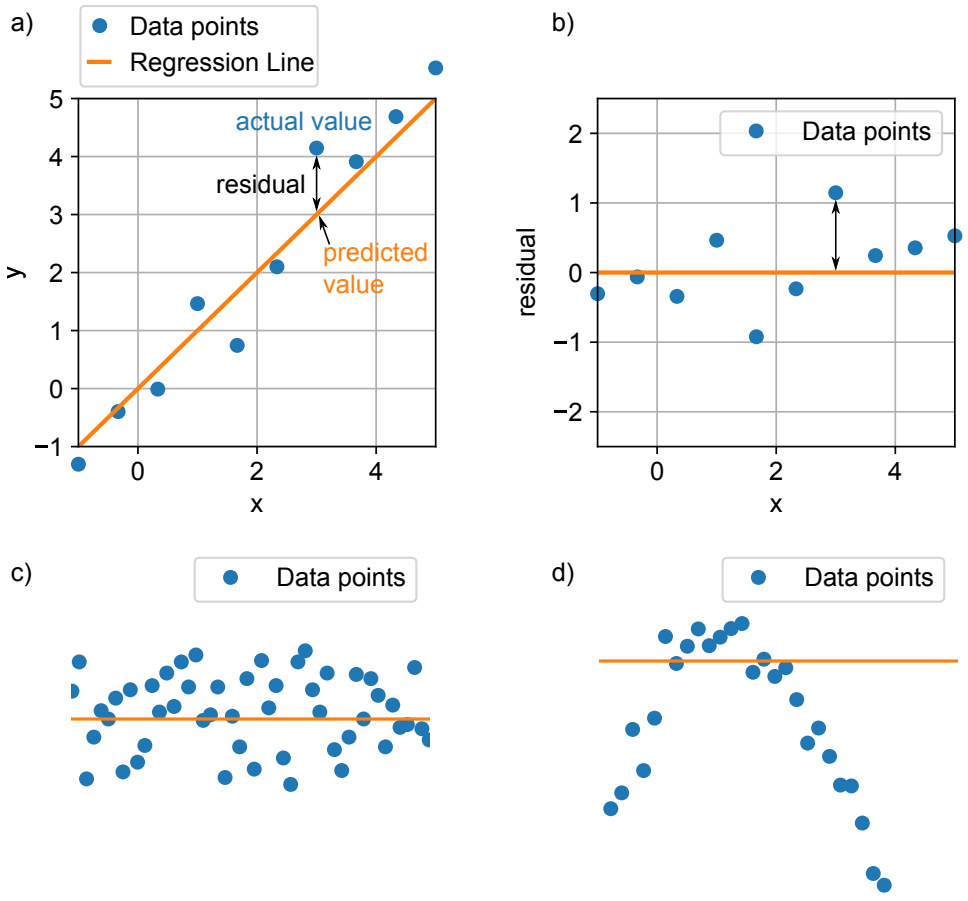

Figure 2: a) Residual on the regression plot and b) corresponding residual plot. Residual plots for good (c) and bad (d) fits.

3.3 Multiple linear regression

In many situations outcomes are influenced not by a single independent variable, but by multiple variables $(x_{\cdot 1}, \ldots, x_{\cdot m})$.
Model class:

$$y = \begin{bmatrix} x_{\cdot 0} & \cdots & x_{\cdot m} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_m \end{bmatrix} + \varepsilon \tag{17}$$

To make the notation simpler one of the variables, e.g., $x_{\cdot 0}$, is assumed to be equal to 1, so that the corresponding parameter θ_0 works out to be the y-offset, or the constant bias term. To simplify the notation when deriving and writing the solution, it is convenient to introduce the following notation that includes the relationship between all the observations in the dataset.

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} x_{10} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots \\ x_{N0} & \cdots & x_{Nm} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{bmatrix} \tag{18}$$

or, more succinctly,

$$Y = X\theta + \varepsilon \tag{19}$$

The model prediction is

$$\hat{Y} = f(X) = X\theta \tag{20}$$

Similar to simple linear regression, we first define the summed square loss function as follows :

$$L(\theta, X, Y) = (Y - \hat{Y})^T (Y - \hat{Y}) \tag{21a}$$
$$= (Y - X\theta)^T (Y - X\theta) \tag{21b}$$

Recall that for matrices A and B, $(AB)^T = B^T A^T$

$$= (Y^T - \theta^T X^T)(Y - X\theta) \tag{21c}$$
$$= Y^T Y - \theta^T X^T Y - Y^T X\theta + \theta^T X^T X\theta \tag{21d}$$

Since $\theta^T X^T Y$ is a scalar, it is equal to its transpose : $\theta^T X^T Y = (\theta^T X^T Y)^T = Y^T X\theta$

$$= Y^T Y - 2\theta^T X^T Y + \theta^T X^T X\theta \tag{21e}$$

The loss function is minimized when $\nabla_\theta L = 0$ This condition gives us

$$\nabla_\theta L(\theta, X, Y) = 0 \tag{22a}$$
$$-2X^T Y + 2X^T X\theta = 0 \tag{22b}$$
$$X^T X\theta = X^T Y \tag{22c}$$

Assuming that the matrix $X^T X$ is invertible, the solution is

$$\hat{\theta} = (X^T X)^{-1} X^T Y \tag{23}$$

3.4 Polynomial regression

What the relationship between the input and output is not linear? A more general class are *polynomial functions*

$$f(x) = \sum_{j=0}^{m} \theta_j x^j \tag{24}$$

To fit a polynomial model to data, one can use a simple trick. Assume that each power of x is another variable, i.e., $x_{ij} = x_i^j$, and apply the multiple regression solution Eq. 23 to fit the coefficients $\hat{\theta}$ of the polynomial function.

3.5 Incremental linear regression

Even though an analytical solution for linear regression exists, we can still choose to apply an incremental algorithm. The loss per item is

$$l(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2 = \left(y_i - x_i^T \theta\right)^2 \tag{25}$$

The gradient of the loss function is

$$\frac{\mathrm{d}}{\mathrm{d}\theta} l(y_i, \hat{y}_i) = -2(y_i - \hat{y}_i) x_i \tag{26}$$

Stochastic gradient descent applied to linear regression therefore yields the update equation

$$\hat{\theta}_i = \hat{\theta}_{i-1} + \eta (y_i - \hat{y}_i) x_i \tag{27}$$

3.6 The curse of dimensionality

If we can calculate multiple linear regression so easily, why do we mostly use few dimensions, i.e., small m? The reason is that the amount of data needed to generalize accurately grows exponentially as the number of features or dimensions grows.

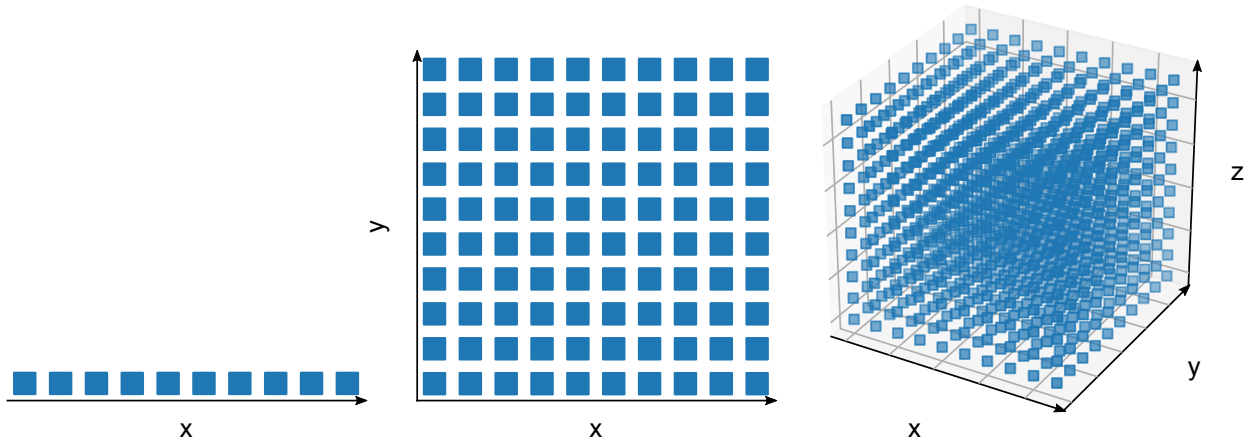

Figure 3: The curse of dimensionality.

The curse of dimensionality arises because the volume of a cube with length a and dimension d scales as a^d. To generalize accurately the model needs to have samples with a certain density. Let's say a point can inform us about its neighborhood up to a certain distance. For instance, in 1 dimensions we might need 10 data points to cover the length a (Fig. 3), then in 2-d we need 10^2 and in d dimensions, we'd need 10^d data points. For example, for a $10x10$ pixel image with 10 grey scale values, we would need 10^{100} data points! For comparison, the total number of atoms in the observable universe is estimated to be "only" 10^{80}.