

# Machine Learning

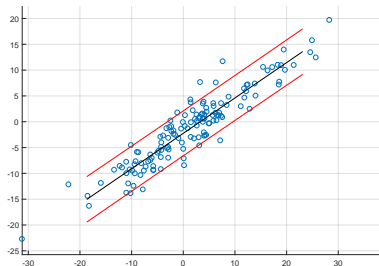
## Supervised learning 2

S. Nõmm

<sup>1</sup>Department of Software Science, Tallinn University of Technology

28.02.2023

# Linear regression: probably the oldest machine learning technique



- Find linear correlation coefficient.
- Compute coefficients of the linear equation

$$\hat{y} = ax + b$$

- Evaluate the model

- In multivariate case it is required to identify coefficients of the model

$$\hat{y} = a_1x_1 + a_2x_2 + \dots + a_nx_n + b.$$

This leads the necessity to choose variables (perform model building).

# Linear regression

- Correlation coefficient.

$$\rho = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

where,  $n$  - is the sample size,  $x$  and  $y$  are the variable of interest.

- $-1 \leq \rho \leq 1$
- Assumption there are exist  $\alpha$  and  $\beta$  such that for any  $i = 1, \dots, n$   $y_i = \alpha x_i + \beta + \varepsilon_i$  holds. Assumption:  $\varepsilon$  is sufficiently small normally distributed.
- The goal of regression is to find estimates of the coefficients  $\alpha$  and  $\beta$ , such that for  $a$  and  $b$

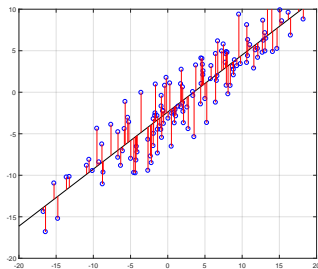
$$y_i = ax_i + b + \hat{\varepsilon}_i$$

sum of squares of  $\hat{\varepsilon}_i$  would be minimal. NB! notation  $\hat{\alpha}$  and  $\hat{\beta}$  is also widely use.

# Least squares method

Least squares method:

$$a = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}; \quad b = \bar{y} - a\bar{x}$$



For an arbitrary number of variables:

$$y = b_1 x_1 + \dots + b_n x_n + b_0$$

then

$$\hat{b} = (X^T X)^{-1} X^T y.$$

where each row of matrix  $X$  is input vector with 1 in the first position.

# Model validation

- Coefficient of determination  $R^2$  and adjusted  $R^2$ .
- Significance of the model and model coefficients.
- Verify assumption that residuals are normally distributed.
- Residual sum squares.  $RSS = \sum_{i=1}^N (y_i - x_i^T \beta)^2$ .
- Sum squares of the regression  $SSR = \sum_{i=1}^N (\hat{y}_i - \bar{y})^2$ .
- Total sum squares or sum of squares about the mean  $SST = \sum_{i=1}^N (y_i - \bar{y})^2$ .
- $R^2$  computed as the ratio of Sum squares of the regression to total sum squares or one minus ratio of Residual sum squares to total sum squares whereas adjusted  $R^2$  is one minus ratio of residual sum squares computed for  $n - 1$  to Total sum squares for  $n - p$  observation points.

# MLE for regression least squares I

- Linear regression is the model of the form

$$p(y|x, \theta) = \mathcal{N}(y|\beta^T x, \sigma^2)$$

where  $\beta$  are the coefficients of the linear model,  $\sigma$  is the standard deviation of  $x$  and  $\theta = (\beta, \sigma^2)$

- Parameter estimation of a statistical model is usually performed by computing MLE  $\hat{\theta} = \arg \max_{\theta} \log p(\mathcal{D}|\theta)$ . remind that  $\mathcal{D}$  denotes the data set

## MLE for regression least squares II

- Assumption: elements of the training set are independent and identically distributed.
- Then log likelihood is given by
$$\ell(\theta) = \log p(\mathcal{D}|\theta) = \sum_{i=1}^N \log p(y_i|x_i, \theta).$$
- As usually instead of maximizing the log-likelihood one may minimize negative log likelihood.
- 

$$\begin{aligned}\ell(\theta) &= \sum_{i=1}^N \log \left[ \left( \frac{1}{2\pi\sigma^2} \right) \exp \left( -\frac{1}{2\sigma^2} (y_i - \beta^T x_i)^2 \right) \right] \\ &= \frac{-1}{2\sigma^2} \text{RSS}(\beta) - \frac{N}{2} \log(2\pi\sigma^2).\end{aligned}$$

# MLE for regression least squares II

- In order to minimize RSS differentiate its equation which lead

$$\nabla\theta = X^T X\beta - X^T y.$$

- Equate it to zero and solve for  $\beta$

$$\beta = (X^T X)^{-1} X^T Y$$

last equation is referred as normal equation.



# Regularization

- Overfitting may be caused by the fact that chosen model structure and data are not conform on another.
- Regularization is the technique used to overcome overfitting.
- Regularization imposes cost or penalty on the cost function and prevent larger values of the coefficients.
- Loosely speaking, regularization shrinks the coefficients towards zero and towards one another.

# Ridge regression

- Ridge regression shrinks the coefficients by penalizing their size.

$$\hat{\beta}^{\text{ridge}} = \operatorname{argmin}_{\beta} \left\{ \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\}$$

$\lambda$  is the nonnegative shrinkage parameter, its large values correspond to the greater amount of shrinkage applied.

- Alternatively the following notation is widely used:

$$\begin{aligned} \hat{\beta}^{\text{ridge}} &= \operatorname{argmin}_{\beta} \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 \\ &\text{subject to } \sum_{j=1}^p \beta_j^2 \leq t \end{aligned}$$

# The Lasso

- Ridge regression shrinks the coefficients by penalizing their size.

$$\hat{\beta}^{\text{lasso}} = \operatorname{argmin}_{\beta} \frac{1}{2} \left\{ \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}$$

$\lambda$  is the nonnegative shrinkage parameter, its large values correspond to the greater amount of shrinkage applied.

- Alternatively the following notation is widely used:

$$\hat{\beta}^{\text{ridge}} = \operatorname{argmin}_{\beta} \sum_{i=1}^N (y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j)^2$$

subject to  $\sum_{j=1}^N |\beta_j| \leq t$

- Computing the lasso solution is a quadratic programming problem.

# Statistical hypothesis testing (brief reminder I)

- Assumption about a parameter of population is a statistical hypothesis.
- Usually a pair of hypothesis is stated  $(H_0, H_1)$ , notation  $(H_0, H_a)$ .
  - ▶  $H_0$  the null hypothesis usually states that there is no statistically significant relationship between two phenomena.
  - ▶  $H_1$  the alternative hypothesis usually states the opposite to the  $H_0$ .
- Choose and compute test statistic and rejection rule.
- Interpret the results.
- What can possibly go wrong?

## Statistical hypothesis testing (brief reminder II)

	Accept $H_0$	Reject $H_0$
$H_0$ is true	Correct	Type 1 Error
$H_0$ is false	Type II Error	Correct

## Model building (feature selection)

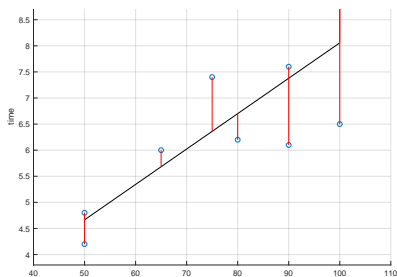
Let us suppose that observed process has  $p$  independent variables  $x_1, \dots, x_p$  and one dependent variable  $y$ . Should one build the regression equation using all  $p$  variables or not?

- Are all the variables  $x_1, \dots, x_p$  uncorrelated?
- Which subset of variables result in a "better" model?
- How to prove that as a result of adding or deleting a variable model quality has improved?

## "Butler tracking company" example

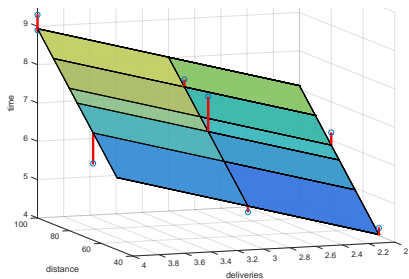
- Independent variables: Distance to drive and number of parcels to deliver. Dependent variable: time.
- Distances to drive for each assignment: 100, 50, 100, 100, 50, 80, 75, 65, 90, 90.
- Number of parcels to deliver: 4, 3, 4, 2, 2, 2, 3, 4, 3, 2
- Time in hours: 9.3, 4.8, 8.9, 6.5, 4.2, 6.2, 7.4, 6, 7.6, 6.1.
- Pearson correlation coefficient between distance and time is 0.81.

## "Butler tracking company" example continued



### Model 1

Is significant  $p = 0.004$ ,  
 $F = 15.1846$  whereas  
 $R^2 = 0.6641$ .



### Model 2

Is significant  $p = 0.000276$ ,  
 $F = 32.9$  whereas adjusted  
 $R^2 = 0.87$ .

Is it enough to say that model 2 is more precise?



## Quality comparison

- To compare different models residual sum of squares (RSS) is used.
- Hypothesis statements:  $H_0 : \text{RSS}_s \leq \text{RSS}_c$   $H_1 : \text{RSS}_s > \text{RSS}_c$ .
- Test statistic (empirical parameter) for ANOVA:

$$F_{stat} = \left( \frac{\text{RSS}_s - \text{RSS}_c}{m} \right) \left( \frac{\text{RSS}_c}{n - p - 1} \right)^{-1}$$

where  $\text{RSS}_c$  is the residuals sum squares of model with more variables,  $\text{RSS}_s$  - is the residuals sum squares of model with less variables,  $m$  number of variables added or removed,  $n$  is the number of observation points,  $p$  - is the number of variables in more complicated model.

- Rejection rule for  $\alpha$  (significance level), degrees of freedom: first is the number of variables added or removed, second is  $n - p - 1$ .
- Decision:
  - ▶ (if adding variables) rejected null hypothesis proves that adding variables caused model quality to increase significantly.
  - ▶ (if deleting variables) rejected alternative hypothesis proves that deleting variables did not cause model quality to significant decrease.

## "Butler tracking company" example continued

- $RSS_1 = 15.8713$ ,  $RSS_2 = 2.2994$  NB! Observe that corresponding MATLAB notation is SSE!!!
- choose  $\alpha = 0.05$  degrees of freedom: first will be 1 (one variable (number of parcels)) were added, second 7 ( $n = 10, p = 2$ ).
- Rejection rule: reject  $H_0$  if  $F_{stat} > 5.5914$
- Compute  $F_{stat} = 17.4411$ . (use table, or MATLAB or EXCEL)
- Reject  $H_0$ . Adding the variable has increased the model quality.

# Linear model building 1

- Choose or determine all the hyperparameters. Possible order limitations, backward elimination / forward selection/ batch processing, set the level of significance and threshold for correlation. These parameters also define stopping criteria.
- Stop when: model is significant, and goodness parameters as expected OR no more variables to add or delete OR maximal or minimal order is reached etc.
- Investigate if available explanatory variables (predictors) are linearly independent. Strong dependencies between variables chosen as "independent" lead problems with inverting matrix  $X$ . Compute multicollinearity matrix where element in  $i$ th row and  $j$ th column is Pearson correlation coefficients computed for variables  $i$  and  $j$ . Based on this table determine subset(s) of variables which are linearly independent.

## Linear model building 2

- Repeat
- Apply mean squares (or other technique) to build the model from selected variables.
- Evaluate significance- and quality- of the model. For quality observe determination coefficient and error. For significance use  $F$  - test and  $t$ -test variable wise.
- If model fail goodness or significance check then return to the previous model and choose another set of variables to add/delete.
- Starting from second iteration prove, using  $F$  - test, that as a result of adding/deliting variables model quality has improved/did not decreased significantly.
- If adding/deliting variables was not successful return to the previous model and if possible chose another variable(s) to add /delete or report the model from previous step.
- If goodness criteria (quality and significance) is met stop and return the model.
- If goodness criteria was not met but adding deleting variables proved to be successful chose the set of variables to be added or deleted ( $t$ -test) on the next step.
- Until stopping criteria is reached.
- Report the results.

## Linear model building 3

Reminder  $p$  - is the number of variables  $n$  is the sample size.

- $F$  -test of overall significance in regression analysis.
- Test for model significance.  $H_0 : b_1 = \dots = b_p = 0$ ,  $H_1 : \exists i : 1 \leq i \leq p \& b_i \neq 0$ .
- Test statistic:

$$F = \frac{\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{p - 1}}{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - p}}$$

- Rejection rule: Determine using F-table or corresponding software function with chosen significance level,  $n$  degrees of freedom in denominator and  $p$  degrees of freedom in nominator.

## Linear model building 4

- $F$  -test to determine significance of change in model quality caused by adding variables

- ▶  $H_0 : RSS_S \leq RSS_C, H_1 : RSS_S > RSS_C.$

- ▶ Test statistic:

$$F = \frac{\frac{RSS_S - RSS_C}{m}}{\frac{RSS_C}{n - p - 1}}$$

- ▶ Rejection rule: Determine using F-table or corresponding software function with chosen significance level,  $n - p - 1$  degrees of freedom in denominator and  $m$  degrees of freedom in nominator.

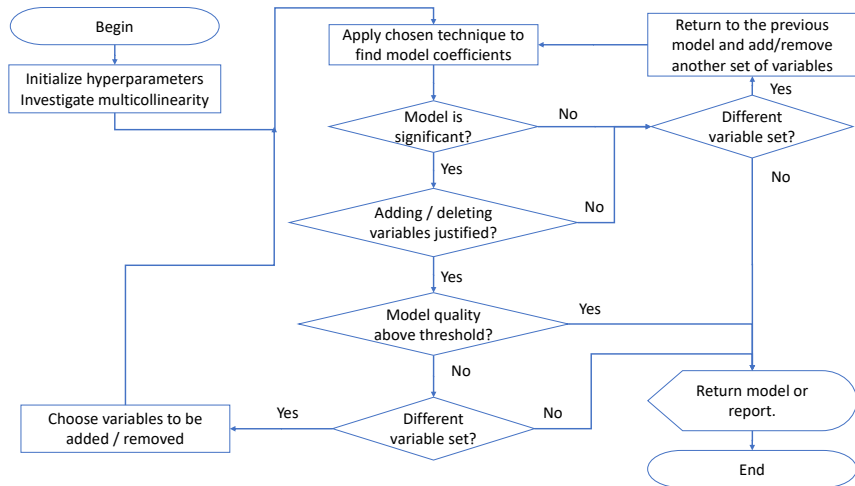
- $t$  - test on individual regression coefficients

- ▶  $H_0 : b_i = 0, H_1 : b_i \neq 0.$

- ▶ Test statistic:  $t = \hat{b}_i / se(\hat{b}_i)$

- ▶ Use  $t$  - table or corresponding function to find rejection rule for chosen significance and  $n - 2$  degrees of freedom.

# Linear model building 5



# Nonlinear regression

- By replacing independent variables  $X$  with a nonlinear mapping  $\phi(X)$ .
- This will lead

$$f_{\theta}(X) = \theta^T \phi(X)$$

- This process is referred as basis function expansion.
- Example: Polynomial regression has basis function  $\phi(X) = [1, x, x^2, \dots, x^d]$ . The model remains linear in the parameters.



# Polynomial regression 1

- Higher degree polynomial models tend to over fit. The coefficients become relatively large, which causes the regression curve to "wobble".
- In order to achieve "encourage" smaller weight values introduce zero-mean Gaussian prior:

$$p(\theta) = \prod_j \mathcal{N}(\theta_j | 0, \tau^2)$$

where  $1/\tau^2$  controls the strength of prior.

- This lead following log-likelihood estimate

$$\ell = \sum_{i=1}^N \log \mathcal{N}(y_i | \theta^T x_i, \sigma^2) + \sum_{j=1}^p \log \mathcal{N}(\theta_j | 0, \tau^2)$$

- The solution is given by:

$$\hat{\theta}_r = (\lambda I + X^T X)^{-1} X^T y$$

# Logistic regression

- Remind that linear regression may be written in the following form:

$$p(y|x, \theta) = \mathcal{N}(y|\mu(x), \sigma^2(x))$$

- This may be generalized to the binary setting as follows:

$$p(y|x, \theta) = \text{Ber}(y|\text{sigm}(\theta^T x))$$

where  $\text{sigm}(\eta) = (1 + e^{-\eta})^{-1}$ . Will be referred as logistic regression.

- Fitting is usually done by maximum likelihood

$$\ell(\theta) = \sum_{i=1}^N \log p(g_i)(x_i|\theta) = \sum_{i=1}^N \left\{ y_i \beta^T x_i - \log(1 + e^{\beta^T x_i}) \right\}$$

- Solving the last one is done by means of iterative algorithm.

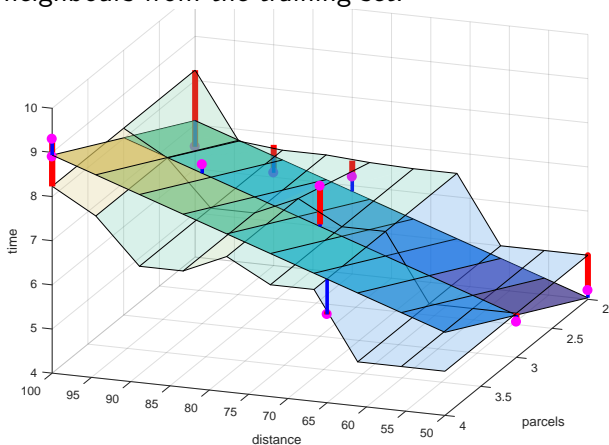
$$b^{\text{new}} = \arg \min_b (z - Xb)^T W (z - Xb)$$

$$z = Xb + W^{-1}(y - p)$$

where  $W$  is a  $N \times N$  diagonal matrix with  $i$ th element  $p(x_i, |b)(1 - p(x_i|b))$

## $k$ -nn regression

The value of the response (dependent variable) defined as the average of its  $k$  nearest neighbours from the training set.



# Regression trees

- Partition the feature space into the set of rectangles.
- Fit a simple model (for example constant) in each rectangle.
- Fitting the model is similar to the case of classification trees.

