Linear regression

Genoveva Vargas-Solar
French Council of Scientific Research, LIG
genoveva.vargas@imag.fr

http://vargas-solar.com/data-centric-smart-everything/

Linear regression assumptions

- Linearity of residuals
- Independence of residuals
- Normal distribution of residuals
- Equal variance of residuals

Linear regression:
linear models (discriminative)

- **Assumptions:**
  - linearity, independence, normal distribution & equal variance of residuals
  - $y|x; \vartheta \sim \mathrm{N}(\mu, \sigma^2)$
- **Normal equations:** X the matrix design, the value of ϑ that **minimizes the cost function** is a closed-form solution such that

$$\boxed{\theta = (X^T X)^{-1} X^T y}$$

  - Design matrix is a matrix of values of explanatory variables of a set of objects.
    - Each row represents an individual object
    - with the successive columns corresponding to the variables and their specific values for that object

Linear regression:
linear models (discriminative)

- A way of calculating the relationship between two variables
  - y dependent, x independent,
  - A and B coefficients determining the slope and intercept of the equation
  - A and B calculated to minimize the error between the models prediction and actual data

$$y = Bx + A \text{ , error} = (\text{Actual} - \text{Prediction})^2$$

$$A = \text{mean}(y) - B\ \text{mean}(x)$$

$$B = \text{correlation}(x,y) \times \text{std}(y)/\text{std}(x)$$

For ZX and ZY standardised versions of x and y, means = z, std = 1

$$ZX_i = [X_i - \text{mean}(X)]/\text{std}(X) \text{ ; } ZY_i = [Y_i - \text{mean}(Y)]/\text{std}(Y)$$

$$r(X,Y) = \text{sum}[ZX_i \times ZY_i] / n - 1 \text{ , } n \text{ –sample size}$$

Linear regression:
linear models (discriminative)

- **Least Means Square algorithm (LMS):**
  - Given the learning rate α, and training set of m data points,
  - the Widrow-Hoff learning rule or LMS, is as follows:

$$\forall j, \quad \theta_j \leftarrow \theta_j + \alpha \sum_{i=1}^{m} \left[y^{(i)} - h_\theta(x^{(i)}) \right] x_j^{(i)}$$

Remark: the update rule is a particular case of the gradient ascent.

- **Locally Weighted Regression (LWR):** is a variant of linear regression that weights each training example in its cost function by $w^{(i)}(x)$, which is defined with parameter $\tau \in \mathbb{R}$ as:

$$w^{(i)}(x) = \exp\left(-\frac{(x^{(i)} - x)^2}{2\tau^2} \right)$$

Regression metrics (1)

- **Basic metrics** — Given a regression model *f*, the following metrics are commonly used to assess the performance of the model:

Total sum of squares	Explained sum of squares	Residual sum of squares
$\mathrm{SS}_{\mathrm{tot}} = \sum_{i=1}^{m} (y_i - \bar{y})^2$	$\mathrm{SS}_{\mathrm{reg}} = \sum_{i=1}^{m} (f(x_i) - \bar{y})^2$	$\mathrm{SS}_{\mathrm{res}} = \sum_{i=1}^{m} (y_i - f(x_i))^2$

- **Coefficient of determination**— r^2, measure of how well the observed outcomes are replicated by the model [0,1]

$$R^2 = 1 - \frac{\mathrm{SS}_{\mathrm{res}}}{\mathrm{SS}_{\mathrm{tot}}}$$

- **Division ratio**— Sum of squared errors $\mathrm{SS}_{\mathrm{res}}$ / *Total sum of squares* SS_{tot}

Regression metrics (2)

- **Main metrics** — commonly used to **assess the performance** of regression models, by taking into account the number of variables *n* that they take into consideration:

Mallow's Cp	AIC	BIC	Adjusted R^2
$\frac{\mathrm{SS}_{\mathrm{res}} + 2(n+1)\widehat{\sigma}^2}{m}$	$2\Big[(n+2) - \log(L)\Big]$	$\log(m)(n+2) - 2\log(L)$	$1 - \frac{(1-R^2)(m-1}{m-n-1}$

where L is the likelihood and $\widehat{\sigma}^2$ is an estimate of the variance associated with each response.

Root Mean Square Error (RMSE) : STD of the residuals, how far are data points from the regression line?

$$RMSE_{fo} = [\Sigma_{i=1}^{N} (Z_{fi} - Z_{oi})2 / N)]^{1/2}$$

$[(f - o)^2]^{1/2}$ f forecasts expected values or unknown results, o observed values

Regression metrics (3)

- **P-value** or calculated probability: when performing a hypothesis in statistics **determines the strength of the result**s
  - Probability of finding the observed or more extreme results when the null hypothesis $\mathbf{H_0}$ of a study question is true
  - **Alternative hypothesis $\mathbf{H_1}$** is there a significant (not due to change) different in blood pressures between groups A and B if A receives the drug and B sugar

p-value [0, 1]	*The claim in trial is called the null hypothesis no difference in blood pressure A,B*
p-value <= 0,05	Strength against the null hypothesis: we can reject the null hypothesis
p-value > 0,05	Accept

Multiple regressions model

- **Assumption**: null hypothesis, multi – collinearity, standard error of coefficients
- **Measures**
  - Global F-Test to see significance of group of independent variables on the dependent variables
  - r^2 / adjusted r^2
  - RMSE, MAPE
  - Residual plot
  - Assumptions of linear regression

Logistic regression
linear models (discriminative)

Dependent variable is binary
Observations are independent of each other
Little or no multicollinearity among the independent variables
Linearity of independent variables and log odds

- **Sigmoid function or** logistic function:

$$\forall z \in \mathbb{R}, \quad \boxed{g(z) = \frac{1}{1+e^{-z}} \in]0,1[}$$

- **Logistic regression:** assume that $y|x; \vartheta \sim$ Bernoulli(φ)

$$\boxed{\phi = p(y=1|x;\theta) = \frac{1}{1+\exp(-\theta^T x)} = g(\theta^T x)}$$

Remark: there is no closed form solution for the case of logistic regressions.

- **Odds ratio** represents the constant effect of predictor X on the likelihood that on output will occur