REGRESSION & CLASSIFICATION

Video Lectures

by Chris Emmery (MSc)

GOALS OF DATA MINING

- Investigating, describing, and sanitizing data.
- Finding patterns in large data sets.
- Through the application and evaluation of algorithms: classification, clustering, regression, rule mining, outlier detection, etc.

THIS LECTURE

Finding patterns through prediction using:

- Regression.
- Classification.

WHAT MAKES PREDICTION POSSIBLE?

Associations between a feature (x) and a target (y).

If:

- Numerical: correlation.
- Categorical: mutual information.

Given X and y $\rightarrow$ supervised learning. Given only X $\rightarrow$ unsupervised learning.

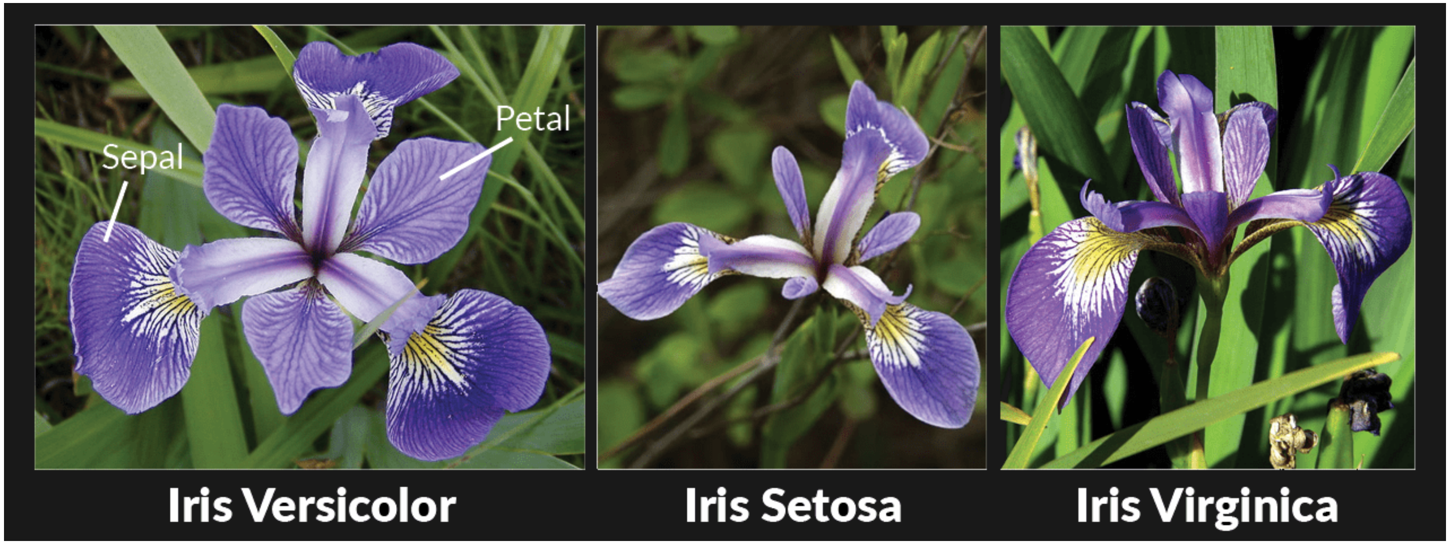
CORRELATION

Pearson correlation:

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

... where n *is sample size,* x *a feature,* y *a target (or feature), indexed by* i*, and* $\bar{x}$ *and* $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} x_i$ *or* y_i *(i.e. the mean).*

IRIS DATASET



IRIS CORRELATION

IRIS CORRELATION II

EXPECTATIONS FROM CORRELATIONS

REGRESSION

CAN WE MODEL THIS AS A FUNCTION?

$f(X) = a \cdot x + b$ or $Y = \beta_0 + \beta_1 \cdot X$

EXAMPLE

city	students (X)	alcohol (Y)	$(X - \bar{X})^2$	$(X - \bar{X}) \cdot (Y - \bar{Y})$
Tilburg	26	41	(26 - 18)^2 = 64	(26 - 18) * (41 - 29) = 96
Eindhoven	21	37	(21 - 18)^2 = 9	(21 - 18) * (37 - 29) = 24
Wageningen	6	9	(06 - 18)^2 = 144	(06 - 18) * (09 - 29) = 240
$\sum$	53	87	217	360

- $\bar{X} = 53/3 \approx 18, \bar{Y} = 87/3 = 29$
- $\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{360}{217} \approx 1.66$
- $\beta_0 = \bar{Y} - \beta_1 \cdot \bar{X} = 29 - 1.66 \cdot 18 = -0.88$
- $\hat{y} = \beta_0 + \beta_1 \cdot X = -0.88 + 1.66 \cdot X$

Sources: RIVM, infogram (2016, 2020)

RESULT

EVALUATION

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} \left(\hat{y}_i - y_i\right)^2}{N}}$$

$$= \sqrt{((42.28 - 41)^2 + (33.98 - 37)^2 + (9.08 - 9)^2)} = 3.28$$

$$R^2 = 1 - = \frac{MSE(f)}{MSE(\mathrm{mean})} = 0.982$$

CLASSIFICATION

REGRESSION VS. CLASSIFICATION

- With regression our y is numerical.
- With classification our y is categorical.

LOGISTIC REGRESSION

$$p(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 \cdot x)}}$$

$$g(p(x)) = \ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 \cdot x$$

LOGISTIC REGRESSION EXPLAINED

k-NEAREST NEIGHBORS

DISTANCES

- Manhattan Distance: $\sum_{i=1}^{n} |x_i - y_i|$
- Euclidean Distance: $\sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$
- Minkowski Distance: $(\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$
- Hamming Distance: (1011101 $\rightarrow$ 1001001 = 2)

More next video!