# Assignment 04

Machine Learning, Summer term 2018

Norman Hendrich, Marc Bestmann, Philipp Ruppel

April 29, 2018

Solutions due by May 07

**Assignment 04.1 (Probabilities, 1+2+2+2+3 points)**

In this exercise, we analyze a simple artificial data-set on vaccination of children. A description of the data is provided in the file vaccination.readme.txt.

a. Read the `vaccination.csv` data into your Python workspace. Determine the numbers of boys/girls, age groups and olderSiblings. Visualize these numbers with bar plots.

b. We are interested in the **marginal probabilities** of individual values in our data. More technically, we are interested in $P(A = a)$, where $a$ is a specific value of a random variable $A$. The random variables correspond to the fields / column names in the data-set, for example, $A =$ gender and $a = 1$ (where 1 denotes "male"). We use short-hand $P(a)$ for $P(A = a)$. $P(a)$ can be estimated from the data using relative frequencies as follows:

$$\hat{P} = \frac{\text{rows with } a}{\text{all rows}}$$

$\hat{P}(a)$ denotes the empirical estimator of $P(a)$ according to the data.

Calculate the empirical probabilities

- to have a vaccination against disease $X$,
- to live on the country side,
- to have at least one older sibling.

c. **Preprocessing** variables can help to better understand the data. A common preprocessing step is to discretize continuous variables. For example, the variable *height* can be transformed into a binary variable *isTallerThan1Meter*.

Calculate the following empirical probabilities:

- What is the probability to be taller than 1 meter?
- What is the probability to be heavier than 40 kg?

Another preprocessing step is the combination of variables. Calculate a variable *diseaseYZ* which denotes whether a child has had either disease $Y$ or $Z$ or both of them. What is $\hat{P}(diseaseYZ)$?

d. **Conditional probabilities** relate two or more variables. $P(a \mid b)$ measures the probability of $a$ given that we know $b$. For example, $P(diseaseX = 1 \mid vacX = 0)$ quantifies the probability that someone has had disease $X$ given that he/she was not vaccinated against $X$.

$P(a \mid b)$ can be estimated using relative frequencies as follows:

$$\hat{P}(a \mid b) = \frac{\text{rows with } a \text{ and } b}{\text{rows with } b}$$

Calculate the following probabilities:

- $\hat{P}(diseaseX \,|\, vacX = 0/1)$
- $\hat{P}(vacX \,|\, diseaseX = 0/1)$
- $\hat{P}(diseaseY \,|\, age = 1/2/3/4)$
- $\hat{P}(vacX \,|\, age = 1/2/3/4)$
- $\hat{P}(knowsToRideABike \,|\, vacX = 0/1)$

where $\hat{P}(a \,|\, b = 0/1)$ is shorthand for $\hat{P}(a = 1 \,|\, b = 0)$ and $\hat{P}(a = 1 \,|\, b = 1)$.

Visualize $\hat{P}(diseaseY \,|\, age = 1/2/3/4)$ and $\hat{P}(vacX \,|\, age = 1/2/3/4)$ as line plots with $age$ on the $x$-axis. What can you conclude from your results?

e. Finally, we take a closer look at the effects of vaccination. Calculate $\hat{P}(diseaseYZ \,|\, vacX = 0/1)$ and compare it to $\hat{P}(diseaseX \,|\, vacX = 0/1)$. What do you conclude from these results? Now, condition additionally on age and calculate $\hat{P}(diseaseYZ \,|\, vacX = 0/1, age = 1/2/3/4)$.

How sure are you that your estimates for $P(diseaseYZ \,|\, vacX = 0/1, age = 1/2/3/4)$ are accurate? What does this depend on?

Plot $\hat{P}(diseaseYZ = 1 \,|\, vacX = 0, age = 1/2/3/4)$ and $\hat{P}(diseaseYZ = 1 \,|\, vacX = 1, age = 1/2/3/4)$ as two lines in one figure with age on the $x$-axis and the probability on the $y$-axis. What do you conclude from your plot?

Remark 1: The effects in (e) due to the confounding variable age are similar to what is known as Simpson paradox. See here: `http://en.wikipedia.org/wiki/Simpson%27s_paradox`.

Remark 2: This artificial data-set was inspired by the KiGGS data-set (`http://www.kiggs-studie.de/english/survey/kiggs-baseline-study.html`). Some people have used this data-set for pro- blematic data analyses to make obscure claims about putative side-effects of vaccination. For an example in German see here: `http://www.efi-online.de/wp-content/uploads/2014/01/UngeimpfteGesuender.pdf`

**Assignment 04.2 (Least-squares regression, 1+2+1+2+1 points)**

In this exercise, you will implement the linear least-squares method for regression. (Note: we have used `polyfit()` already, but here you are building the matrix and solving the equations yourself...).

a. Data preparation.

- Load `reg1d.mat`. Plot training and test data.
- Preprocess the training data by concatenating 1 (for the bias term) to each training point.

b. Learning

- Write a function `least_squares(X, Y)` which computes the weight vector of the least-squared solution for input points $X \in \mathbb{R}^{n \times d}$ with target values $Y \in \mathbb{R}^{n \times 1}$, where $n$ denotes the number of points and $d$ is the number of features (dimensions per point).
- Calculate $w$ using `least_squares(X, Y)` for the given training data. Plot the prediction of the resulting classifier into your previous plot.

c. Evaluation

- Write a function `err = lossL2(Y, Y_pred)` which returns the empirical squared error of predicting `Y_pred` instead of $Y$.
- What is the average L2 loss of the classifier on the test data?

d. Non-linear features

  - Add quadratic and cubic basis functions to your input features (add new columns for $x^2$ and $x^3$ in addition to $x$ and 1).
  - Re-run learning and evaluation.

e. Outlier

  - Add an extreme outlier to your training data:

```
X_train = numpy.append( X_train, 1.05 )
Y_train = numpy.append( Y_train, -10 )
```

  - Run your code to see its effect on linear least-squares regression.