Final Project – Basic Probability: Programming

Fall 2016, Master of Logic, University of Amsterdam

Instructors: Christian Schaffner and Bas Cornelissen

Submission deadline: Friday, 23 December 2016, 8 p.m.
Assessment deadline: Friday, 13 January 2017, 8 p.m.

1 Project Description

In this project you will perform your own data analysis in groups of 2-3 people. Your task is to use a technique known as linear regression (see below) to predict the median price of houses in suburbs of Boston in the 1980s. The data set and all relevant information about it can be found here. The variable that you need to predict is stored in the last column.

This project will teach you two things: first, it introduces you to a continuous distribution, namely the Gaussian. Linear regression and other Gaussian models are very useful in practice and you will see them a lot when working with data. Second, you will have to interpret your results. Rather than just running the algorithm, you will have to make sense of the algorithm's output. In particular, you will express what the weights learned by linear regression tell you about patterns in the data.

2 Linear Regression

To familiarise yourself with linear regression and its implementation, watch the videos of Sections 2 and 4 of Andrew Ng's machine learning course. You can also sign up for it on coursera to see them in higher resolution. His explanations are great, but he does not really touch on the math of linear regression.

From a probabilistic view point, we assume that our data are independently distributed. This is a weaker assumption than we have made so far because we do not assume that the distributions of our data points are identical! In fact, linear regression is all about working with a different distribution for each data point.

In linear regression we assume that each data point x_i, $(1 \leq i \leq n)$ follows a Gaussian or normal distribution. The probability density function

of this distribution is

$$P(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) . \tag{1}$$

What is special about the normal distribution is that its mean and variance are equal to its parameters. In particular, we have $\mathbb{E}[X] = \mu$ and $\mathrm{var}(X) = \sigma^2$ for $X \sim \mathcal{N}(\mu, \sigma^2)$.

Linear regression assumes that the variance parameter σ^2 is shared by all distributions but that each data point x_i was generated from a Gaussian with mean μ_i. This mean is computed as $\mu_i = w^\top h(x_i)$ where w is a weight vector and $h(x_i)$ is a vector of predictors[1] of x_i. The mean is thus a linear combination of the data point's predictors. Each coefficient in w relates its predictor linearly to the output variable.

3 Your Task

Your task is to implement linear regression for the Boston housing data set. Every group of students first implements a baseline regression model and report results on that. A baseline of this study is the most basic implementation of the linear regression method on this dataset (you have to justify your choice of baseline!). Thereafter, you try to improve your model and get better predictions. You then write a report, detailing your improvements and describing what you have learned about the data set by running linear regression.

4 Evaluating Linear Regression

Since linear regression makes continuous predictions, it cannot simply be evaluated on a wrong/correct basis. The error of linear regression is induced by the distances of the predictions to the true values. These distances are known as residuals. The sum of their squares is the (squared) error[2]. Take a look at the plots produced during the in-class exercises of week 6. If the line was the output of a linear regression model, the residuals would be the vertical distances of the points to the line.

Since the residual error grows as a function of the number of data points, it is not a meaningful way of evaluating a regression model. However, the residuals are again normally distributed. Standardly, it is assumed that

[1] For the sake of explanation, we make a strict difference between features and predictors. Features are attributes of a data point whereas predictors are values supplied to the model.

[2] Squaring is necessary here because the negative and positive residuals would cancel each other out otherwise.

$\mu = 0$ for that distribution.[3] If the variance of their distribution is small, the regression line provides a tight approximation to the true values. If the variance of that distribution is large, the approximation is pretty shabby.

The standard quantity for assessing linear regression models is the R^2 which can be computed as

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(x_i - w^\top h(x_i)\right)^2}{\sum_{i=1}^{n} \left(x_i - \frac{1}{n}\sum_{j=1}^{n} x_j\right)^2} \,. \tag{2}$$

Notice that the numerator is the sum of the residual errors and the denominator is the sample variances of the data. The R^2 thus expresses what fraction of the observed variance in the data is captured by the model. The R^2 has the convenient property of being standardized, i.e. to be bounded by 0 and 1. The better the linear regression fits the data, the closer the value of R^2 is to 1.

5 Types of Features and interpretation

Roughly speaking there are two types of features: continuous ones and categorical ones.

Continuous Features Let us assume that our model only has one continuous feature f with coefficient α. This means that for each point increase in f, the predicted value experiences a change of α. If α was -2, the predicted value would decrease by two for each point increase of f. This is a general property of continuous features in linear regression: their values are linearly related to the output value.

Categorical Features Categorical features are all features for which the assumption that their values are linearly related to the output does not make sense. Consider a feature *gender*. First of all, it needs to be turned into a number to be useful for regression. So let us define the mapping male $\mapsto 0$, female $\mapsto 1$.[4] Then we would certainly not want to assume that the values of *gender* are linearly related to whatever output we are dealing with. Rather, it is the level[5] *female* that causes a change in the output.

Whenever we are dealing with categorical features, we need to define a standard or default level. This level can be chosen arbitrarily (in the example above it was chosen to be male). All coefficients of the other levels then

[3] This assumption is made because residuals can be both positive and negative and there is no reason to assume that residuals should be "more positive" or "more negative" on average.

[4] This mapping is arbitrary in that we might also map males to 1 and females to 0.

[5] The values of categorical features are often called *levels*.

have to be interpreted with respect to that default. To make this concrete, let us assume that we model football-playing ability and that our categorical features is *nationality* with levels Holland, Germany and Switzerland. We arbitrarily chose Holland as default. We then introduce two(!) predictors into the model, namely Germany and Switzerland. Each of these predictors is either 1 (when the nationality matches the level) or 0 (otherwise). After training our model, we get a coefficient of 5 for Germany and -3 for Switzerland. The way to interpret these coefficients is that being German increases your ability in football by 5 compared to being Dutch whereas being Swiss decreases it by 3 (again compared to being Dutch).

The reason that we need to introduce two predictors when dealing with 3 levels is that if we assigned the numbers 1 and 2 to Germany and Switzerland, we would again assume a linear relationship based on nationality. In general, introducing a predictor for each non-standard level allows us to estimate a separate coefficient per level.

A pre-final remark: it is not always clear whether a feature should be seen as continuous or categorical. In the housing data, we have the RAD feature. Should that be continuous or categorical? This is something you may want to experiment with.

Cautionary Note Shockingly often, people (in particular researchers) try to ascribe a causal effect to predictors. This is not warranted by the model! In the football example above, being German goes hand in hand with having a higher football ability than being Dutch. This is simply an observation, however. It does not mean that being German *causes* you to be a better footballer. Rather, it means on average Germans are better footballers than Dutch (for reasons that may or may not be related to nationality). You should thus be careful when describing and interpreting your results.

6 Learning

Learning a linear regression model can be done with the technique of least squares. This basically means that you try to iteratively minimise the sum of squares of the residuals. The details are explained in the videos.

7 Improving your Algorithm

There are several ways to improve your algorithm. As discussed in the videos, you may use regularisation. You can also try to build new predictors by transforming or combining features. All of this is up to you. The important thing is that in the end you can show a real improvement through an increase in R^2.

8 Report

Each group writes a report of at most 4 pages (not including the space taken up by references). All reports must be created using the EACL-2017 LaTeX template and follow the instructions provided therein. Your report must contain the following sections:

- **Authors** Specify who contributed to the final project.
- **Abstract** A short summary of your work and your findings. Use the abstract environment!
- **Introduction:** Formally describe linear regression and set out the goal of your data analysis.
- **Improvements:** Give a description of the improvements that you made. You do not need to list everything you have tried here. Rather, focus on the things that worked and got you to your final results.
- **Experiments:** Shortly describe the specifics of the data set and then report your results for the baseline model and possible extensions. Your results need to be supported by tables and/or graphs. If you made several extensions, it is good practice to report results as you add in extensions one by one. This way, one can better determine the contribution of each individual improvement.
- **Conclusion:** Describe what you have learned from your analysis about the data set. Also state whether you think that your results are satisfactory and reliable.

9 Deliverables

In order to complete this project you need to submit a .zip file containing (possibly subfolders with) the following items:

- **Report:** A report as described above.
- **Code:** Any code that you have written. This code should be executable either from the command line or from Pycharm (or both).
- **Readme:** A text file describing the contents of the folder. This file should also contain a link to the data set which you used. Most importantly it should contain very detailed instructions on how to run your code!

10 Assessment

We again have a peer-review phase (until January 13, 2017) where your group gets to assess the work of two other groups, as well as your own. Besides written feedback, you can suggest a final grade.

The lecturers of the course determine the final grade for the project based on these self-/peer-assessments and their own judgement. Every student in a given group receives the same grade.

For the assessment, take into account the following aspects:

- Editorial quality of the report: Has the required LaTeX template been used? Are all points outlined in Section 8 of the project description addressed properly? Is the report easy to read? Does it contain graphs and tables? Are there proper references?
- Code quality: Is the code executable? Does the readme file explain how to run it? Are there doc-strings in the code? Is it sufficiently commented so that you can easily follow what happens?
- Scientific results: Is there a proper baseline implementation? Is it clear which extra steps have been taken? Did they pay off? Are these improvements properly explained? How do the obtained results compare to your own? Are the stated conclusions about the dataset and method valid?
- General comments and suggestions for improvements.
- Suggestion for a grade.