IST Austria: Data Science and Scientific Computing
Track Core Course 2016/17 – Segment 1: "Predictive Models"

Instructor: Christoph Lampert <chl@ist.ac.at>
TAs: Bor Kavcic <bor.kavcic@ist.ac.at>, Katharina Ölsböck <katharina.oelsboeck@ist.ac.at>,

Exercise Sheet 2

1) Literature

a) Read this paper: [A. Halevy, P. Norvig, F. Pereira, *"The Unreasonable Effectiveness of Data"*, IEEE Intelligent Systems, vol. 24, no. 2, pp. 8-12, March/April, 2009] What is the author's main message in a single sentence?

b) Read Chris Anderson's `wired` article: "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete". What is his main message in a single sentence?

c) Anderson quotes Norvig with the statement "All models are wrong, and increasingly you can succeed without them." (which Norvig claims he never said). Do you agree?

d) Visit `http://karpathy.github.io/2015/05/21/rnn-effectiveness/` and read Andrej Karpathy's blog post about Recurrent Neural Networks. You can skip the technical parts, but have a look at the example network outputs when trained on Shakepeare's writing, Wikipedia, LaTeXor the Linux kernel.
Which one is your favorite example and why?

2) Hands-On

a) Adjust your implementation of ridge regression from the first exercise such that it automatically uses the alternative (dual) representation when the data dimension is higher than the number of samples.

b) Write a routine that performs kernel ridge regression for an arbitrary kernel function (you can make the kernel function an argument, or have it defined externally).

c) Write a model selection routine that uses a validation set to choose between the following kernels

- linear kernel, $k(x, x') = x^\top x'$
- linear kernel with bias term, $k(x, x') = x^\top x' + 1$
- RBF kernel with bandwidth, $k(x, x') = e^{-\gamma\|x-x'\|^2}$, with $\gamma \in \{0.1, 1, 10\}$
- polynomial kernel with degree $k(x, x') = (1 + x^\top x')^p$, for $p \in \{2, 3, 4\}$

d) Apply c) to the `diabetes` dataset. Which kernel is selected?

3) More Hands-On Experience with Text

We will try to predict how popular tweets are based on their text.

a) Download the JSON file `http://www.ist.ac.at/~chl/courses/DSSC-TCC17/trump-2014.json` (4MB). It contain all tweets by Donald Trump from the year 2014. If your language of choice cannot handle JSON files, then use the CSV file from same URL with 'json' replaced by 'csv'.

b) From each tweet, extract the entries `retweet_count` and `text`. Note: some tweets do not have a `retweet_count` entry. Discard those.

c) For each tweet, compute 1) a target variable $y = \log(1 + \texttt{retweet\_count})$; 2) a vector representation x for the text by extracting all words and averaging the corresponding `glove.twitter.27B.25d.txt` word vectors (see the previous exercise sheet). For this, remove all special characters, punctuation etc, and convert words to lowercase. Ignore words for which the text-to-vector does not have an entry. If a tweet has no usable text, assign it the vector of all zeros.

d) Train a linear regression model with bias term, $f(v) = w^\top v + b$, on the data.

e) After training the model, download the JSON file `http://www.ist.ac.at/~chl/courses/DSSC-TCC17/trump-2013.json` (5MB) and process it in the same way as above (or use the CSV). Evaluate your regression model by computing the correlation coefficient of the predictions to the true values, and by plotting the predicted number of retweets against the true number of retweets.

f) For each word that appears in the training tweets and that is in the dictionary, compute the prediction score. Which 20 words get the most positive and which get the most negative scores?

4) Additional exercises (voluntary, to get more practice)

a) Download the JSON file `http://www.ist.ac.at/~chl/courses/DSSC-TCC17/trump-2015.json` and process it in the same way as above. Evaluate your regression model from 3) by plotting the predicted number of retweets against the true number of retweets. There's a good chance the results are worse than for 2014. Can you find out why that is the case?

b) Download the JSON file `http://www.ist.ac.at/~chl/courses/DSSC-TCC17/trump-2016.json`, train a new regression model and and redo the analysis of 3f). What do you observe?

c) Instead of averaging the vectors of all words in 3c), try different text representations, e.g. removing stop words (stop word lists are available online) or weighting the vectors like *tf-idf* does.