Homework 3: the Bordeaux wine equation

Fundamentals of Data Science, 2016/2017

In the following, "ID" denotes your student's ID if you have one, else it denotes your last name (please, suppress accents if present).

The goal of this homework is to submit a Python module and a Python script, named "ID-lib.py" and "ID-run.py" respectively. For example, if your student ID is 12345, the two files will be named `12345-lib.py` and `12345-run.py`.

ID-lib.py

The module "ID-lib.py" must implement the following functions.

```
descent(y, x, alpha = 1e-3, itr = 1e2, eps = 1e-6)
```

This function receives a numpy array `y` of length m and a numpy array `x` of size $m \times p$. It then performs a gradient descent on $f(\Theta) = \sum_{k=1}^{m}(y_k - \Theta^T x_k)^2$, where Θ is a vector of length p and y_k and x_k are respectively the k-th entry of y and the k-th row of x. You can set initially $\Theta = \mathbf{0}$. The descent has learning rate `alpha`. The descent stops as soon as (a) `itr` update iterations have been done, or (b) the relative variation $|\Theta^{(i+1)} - \Theta^{(i)}|/|\Theta^{(i)}|$ between two successive iterations, where $|\cdot|$ denotes the 1-norm, is not larger than `eps`. The function returns the final value of the vector Θ as a numpy array.

```
r2(y, c, x)
```

This function receives three numpy arrays: `y` of length m, `c` of length p, and `x` of size $m \times p$. It then returns the coefficient of determination, R^2, of the linear model $f(x_k) = c^T x_k$ on the data points $\{(x_k, y_k)\}_{k=1,\dots,m}$, where y_k and x_k denote respectively the k-th entry of y and the k-th row of x.

ID-run.py

The script "ID-run.py" must load the dataset `wine.csv` and

1. Normalize each feature of the dataset (except for `Price`) so to have mean 0 and standard deviation 1 across the observations. **Note: if you do not normalize, you might experience troubling numerical issues.**
2. Perform a linear regression on the model `Price` $= a_0 + a_1 \cdot$ `AGST`, then print the coefficients and the R^2 of the model.
3. Produce a scatter plot of `Price` against `AGST`, together with the regression line; the script must save the plot to a file named `ID.png`, in PNG format.

4. Perform a linear regression on the model $\texttt{Price} = a_0 + a_1 \cdot \texttt{AGST} + a_2 \cdot \texttt{WinterRain} + a_3 \cdot \texttt{HarvestRain} + a_4 \cdot \texttt{Age}$, then print out the coefficients and the R^2 of the model.

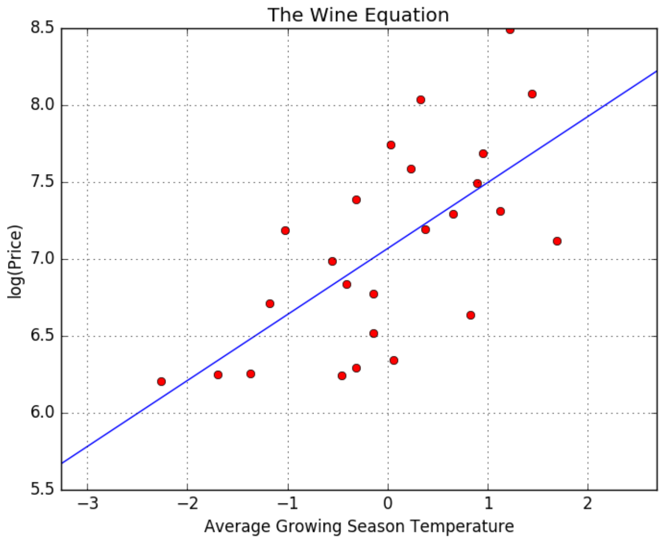
Remark 1: for each regression you'll have to try different values of `alpha`, `itr` and `eps` until you find that the descent converges (and R^2 takes on reasonable values).
Remark 2: I will launch "ID-run.py" with the command `ipython ID-run.py`, using Python 3.x. Make sure you are not relying on features specific of Python 2.x.
Remark 3: your script must take no more than 30 seconds, after that it will be killed.

Submitting the homework

You must send the two scripts by e-mail, not later than Thursday November 6 at 23:59, to the address `fds2016lab@gmail.com`. **The subject of the e-mail must be "ID hw3" (take care of the space).**

Sample output

```
$ timeout 30 ipython 12345-run.py
[ 1.23456789  1.23456789]
R2 = 0.333333333333
[ 1.23456789 -1.23456789  1.23456789 -1.23456789  1.23456789]
R2 = 0.666666666666
```


The Wine Equation
log(Price)
Average Growing Season Temperature