End-to-End Machine Learning Project

* Popular open data repositories
  + UC Irvine Machine Learning Repository
  + Kaggle datasets
  + Amazon’s AWS datasets
  + Meta portals (they list open data repositories)
  + Data Portals
  + OpenDataMonitor
  + Quandl
  + Wikipedia’s list of Machine Learning datasets
  + Quora.com
  + The datasets subreddit
* Data processing components is a pipeline (they are asynchronous)
* V common since data is large so large pipelines since large transformations

Data store

Another data transformation

Data transformation

A simple pipeline

* Each component pulls in a large amount of data, processes it, and spits out the result in another data store
* Then, some time later, the next component in the pipeline pulls this data and spits out its own output. Each component is fairly self-contained: the interface between components is simply the data store. This makes the system simple to grasp (with the help of a data flow graph), and different teams can focus on different components. Moreover, if a com‐ ponent breaks down, the downstream components can often continue to run nor‐ mally (at least for a while) by just using the last output from the broken component. This makes the architecture quite robust. On the other hand, a broken component can go unnoticed for some time if proper monitoring is not implemented. The data gets stale and the overall system’s performance drops.
* Your next step is to select a performance measure. A typical performance measure for regression problems is the Root Mean Square Error (RMSE). (statquest neural network)
* Some terms:-
  + M= size
  + x (i)=value for ith term
  + y(i)= output for x(i)
  + X=superset of all x(i) with values being the transpose rep as x(i)T
  + H= hypothesis or the prediction function
  + ŷ (i)= h(x(i)) (y hat is the predicted value)
  + RMSE(X,h) is the cost function measured on the set of examples using your hypothesis h.
  + We use lowercase italic font for scalar values (such as m or y (i) ) and function names (such as h), lowercase bold font for vectors (such as x (i) ), and uppercase bold font for matrices (such as X)
* Rmse is not omnipotent eg if n(outliers)>>> mean abs error to be used
* take a look at the top five rows using the DataFrame’s head()
* The info() method is useful to get a quick description of the data, in particular the total number of rows, each attribute’s type, and the number of nonnull values
* You can find out what cate‐ gories exist and how many districts belong to each category by using the value\_counts()
* The describe() method shows a summary of the numerical attributes
* you can call the hist() method on the whole dataset (as shown in the following code example), and it will plot a histogram for each numerical attribute
* Scikit-Learn provides a few functions to split datasets into multiple subsets in various ways. The simplest function is train\_test\_split(), which does pretty much the same thing as the function split\_train\_test(), with a couple of additional features. First, there is a random\_state parameter that allows you to set the random generator seed. Second, you can pass it multiple datasets with an identical number of rows, and it will split them on the same indices
* Selecting test cases which represent the whole population called stratified sampling: the population is divided into homogeneous subgroups called strata, and the right number of instances are sampled from each stratum to guarantee that the test set is representative of the overall population.
* If random sampling was used 12% chance that result will be skewed