CMSC 33750:
Machine Learning in Cancer

Programming Assignment 1

October 9, 2017
Version 1.0

This programming assignment uses data from the NCI Genomic Data Commons. You can download the data using the curl command at the end of this assignment.

Each record is a JSON structure associated with a case, and contains: the case number; the primary site; and the chromosome, the start position, and the stop position of the mutation; and some additional information. Please see below.

The goal of this programming assignment is to write a classifier that given some mutational information about a case predicts the primary site associated with the cancer. The point of this assignment is simply to become familiar with the GDC, its data, and to build a basic classifier and to evaluate its performance.

Please carefully write up your assignment, explaining:

1. How you prepared the data for modeling, including the attributes of the dataset you prepared after extracting and cleaning the data.
2. What features you used and how they were computed.
3. How you split the dataset into a training set and validation set.
4. How you built the classifier.
5. How you evaluated the performance of the classifier.
6. What suggestions that you have for improving the classifier.

Please work individually.

It may be useful to find the closest gene to each SNP returned by the query. One way to do this is to use the closest command from bedtools (https://github.com/arq5x/bedtools2/).

You can find a BED file here:

https://www.dropbox.com/s/vd6is98wo7fojlk/gencode.v22.annotation.bed.gz?dl=0

Please prepare:

1. a written report addressing questions 1-6 above.

2. The code you wrote to clean the data, build the features, build the classifier, train the classifier, and evaluate the classifier.

The assignment must be turned in by midnight, **October 25, 2017**.

Additional information will be provided about how to turn in the assignment.

JSON Structure:


```
{
      "ssm": {
        "mutation_subtype": "Single base substitution",
        "end_position": 11394531,
        "start_position": 11394531,
        "genomic_dna_change": "chr12:g.11394531C>A",
        "mutation_type": "Simple Somatic Mutation",
        "chromosome": "chr12"
       },
       "case": {
        "project": {
          "project_id": "TCGA-COAD"
        },
        "primary_site": "Colorectal"
       },
       "id": "317e2e60-8198-5620-bac8-b7e4afc6ae5d"
     },
```


Getting the Data: Here is a query to retrieve data from the GDC:


```
curl
'https://api.gdc.cancer.gov/ssm_occurrences?fields=case.project.project_id,case.primary_site,ssm.muta
tion_type,ssm.start_position,ssm.end_position,ssm.mutation_subtype,ssm.chromosome,ssm.genomic_
dna_change&pretty=true&filters=%7b++%0d%0a+++%22op%22%3a%22AND%22%2c%0d%0a+++%22co
ntent%22%3a%5b++%0d%0a++++++%7b++%0d%0a+++++++++%22op%22%3a%22in%22%2c%0d%0a+++
++++++%22content%22%3a%7b++%0d%0a++++++++++++%22field%22%3a%22ssm.consequence.transcr
ipt.annotation.impact%22%2c%0d%0a++++++++++++%22value%22%3a%5b++%0d%0a++++++++++++++
+%22HIGH%22%0d%0a++++++++++++%5d%0d%0a+++++++++%7d%0d%0a++++++%7d%2c%0d%0a++++
++%7b++%0d%0a+++++++++%22op%22%3a%22in%22%2c%0d%0a+++++++++%22content%22%3a%7b+
+%0d%0a++++++++++++%22field%22%3a%22ssm.mutation_subtype%22%2c%0d%0a++++++++++++%22
value%22%3a%5b++%0d%0a++++++++++++++++%22Single+base+substitution%22%0d%0a++++++++++++
%5d%0d%0a+++++++++%7d%0d%0a++++++%7d%0d%0a+++%5d%0d%0a%7d&size=200000' >
GDC_muts.json
```

Computing the nearest gene: It may be useful to find the closest gene to each SNP returned by the query. One way to do this is to use the closest command from bedtools (https://github.com/arq5x/bedtools2/).

You can find a BED file here:

https://www.dropbox.com/s/vd6is98wo7fojlk/gencode.v22.annotation.bed.gz?dl=0