W7 Assignment – Machine Learning 2

Computational Linguistics

John Gamboa

April 22, 2022

1 Dataset Splitting

1.1 No split

Consider the following situation. You are given a dataset D containing data divided into 3 classes c_1, c_2 and c_3. Your goal is to create a Nearest Neighbor classifier ($k = 1$) that is able to generalize to new data. For data, when training this classifier, you give **your whole dataset** D.

Then you wonder how well this dataset generalizes to data it has never seen. For that, you make a subset of D and use it as a training set. What would be the accuracy of your classifier on this subset of D? (select the correct answer)

- Something close to 100%, but not 100%
- You have no way to know
- 0%
- 100%
- 33%

`answer:` ________________________________

1.2 Goals

When splitting your dataset, there is a number of things you would like your split to consider. Ideally, you'd like your split to be such that...

(choose *TRUE* or *FALSE*)

a) your test set is representative and balanced, so that you can test all "corner cases" in the data

answer: ________________________________

b) your test set as has as much data as possible, so that you can be very sure about how well the learnt model generalizes to new data

answer: ________________________________

c) your training set is bigger than the test set and validation set

answer: ________________________________

d) your training set homogeneously represents only certain elements of the population, to make sure it doesn't learn wrong cases

answer: ________________________________

e) your validation set contains data points in common with the test set

answer: ________________________________

f) your validation set is bigger than the training set

answer: ________________________________

g) your test set contains data points in common with the training set

answer: ________________________________

h) your validation set is representative and balanced, so that you, when tuning your hyperparameters, you can make sure that they consider all corner cases in your data

answer: ________________________________

i) your validation set has as much data as possible, so that your hyperparameters can get tuned to the best values possible

answer: ________________________________

j) your test set contains only the "corner cases" (and doesn't have data related to the most common cases), because it is on the "corner cases" that you are interested in

answer: ________________________________

k) your test set is bigger than the training set

answer: ________________________________

l) your training set has as much data as possible, so that your algorithm can learn well

answer: ________________________________

m) your test set represents a different population than the training set

answer: ________________________________

n) your training set is representative and balanced, so that it can learn about all ”corner cases” in the data

answer: ________________________________

2 Regression

2.1 Theory

Let’s say you are given a dataset D, which contains several samples d_i, each of which is represented as a point in a graph.

A note on the jargon

The “columns” of each sample d_i (i.e., the coordinates that determine the position of the point d_i) are generally referred to as its “features”. E.g., if your samples are measurements of patients, then the “length of the arm” and “size of the eye” of each patient could be “features” inside each sample d_i.

Choose the correct answer:

- When performing regression, you normally use exactly one feature to describe exactly one other feature. I.e., you cannot combine a number of features.
- When performing linear regression, your goal is generally to find a line that best describes the relationship between a number of the features in your data.
- None of the other alternatives is correct

answer: ________________________________

2.2 Coding

Assume you are given a dataset D which is divided into “input features” X and “output features” Y. In other words, you could divide your dataset into $D = (X, Y)$, where X are the features are your input data points, and your goal is to predict the values in the Y.

Let’s say that your goal was to define a line $y = Ax + b$ for all elements $y \in Y$ and $x \in X$ (as we’ve done in the class). For this, you decided you’d use `sklearn`, and created the following object

```
import sklearn
from sklearn import linear_model
reg = linear_model.LinearRegression()
```

Write the line of code that you would run so that you ”linear regressor” would learn the coefficients `A` and `b` of this line.

code: ____

3 Gradient Descent

3.1 Limitations

When performing gradient descent, you normally have a number of parameters that you can vary (in our case, A and b) and you are given a function L that is usually referred to as the “loss” or “cost”. Your goal is to minimize L, i.e., to find the values of A and b that will cause L to be the least possible. What limitations have we seen about it? (choose the correct alternative)

- It can only work with losses that look like straight lines.
- It cannot work with losses that look like a “U”.

- If the loss is varies a lot for different values of the parameters, producing lots of small “U’s” in different points, then the Gradient Descent may end up stuck in one of these “U’s”
- It only works with losses that look like a “U” and never produces results otherwise

`answer:` ______________________________