Machine Learning Assignment 54

George Meza

March 2, 2021

Problem 1

(1.a)

Solution

1. 0.04690415759823429
2. 0.037416573867739396,
3. 0.061644140029689765
4. 0.1224744871391589
5. 0.15811388300841897
6. 0.158113883008419
7. 0.08831760866327845
8. 0.24494897427831783
9. 0.21213203435596428
10. 0.18708286933869708
11. 0.3959797974644666
12. 0.17320508075688776
13. 0.22759613353482086

(1.b)

Solution

The five closest neighbors ID's and cookie classification were 2(Shortbread), 1(Shortbread), 3(Shortbread), 7(Sugar), and 4(Shortbread)

(1.c)

Solution

The classification that appeared the most were the shortbread cookies, with 4 of the 5 closest neighbors being them, and from this we can infer that the recipe would most likely be a shortbread recipe.

(2.a)

Solution

If we use 13 cookies(the whole data set), we get fortune cookies as the most cookies with the closest distances and this is dangerous because the only reason fortune cookie wins is because there are more fortune cookies in the data set than any other cookie.

, 0.0671098396587056, 0.16237358723960857 **(2.b)**

Solution

For fortune we get 0.23919918309016733, for shortbread we get 0.0671098396587056, and for sugar we get 0.16237358723960857. This solves the problem in part a because now we know which cookie had the closest distance on average with each type of cookie and we get shortbread as the answer.

Problem 2

(a)

['Name', 'ACT', Extracurricular', 'Accept/Reject']
['Martha', '33', '1', 'Accept']
['Jeremy', '34', '0', 'Reject']
['Alphie', '35', '1', 'Accept']
['Dennis', '30', '1', 'Reject']
['Jennifer', '36', '1', 'Accept']
['Martin', '29', '1', 'Reject']
['Mary', '36', '1', 'Accept']
['Dean', '31', '1', 'Reject']
['Adam', '36', '0', 'Accept']
['Jeremy', '32', '0', 'Reject']

(b)

A logistic model will be used, going from 0.001 to 0.999. Our terms will be ACT score, whether they did an extracurricular or not and whether or not they got accepted. We also have one interaction term between the ACT and the extracurricular. The model goes from 0.001 to 0.95 because there is a cutoff between whether or not you can get in based on these 2 factors, i ignored percentile because it had a very little impact on the acceptance rate. A kid was in the 80th percentile and got in but a kid with a 95 percentile couldn't. Percentile would also depend on the type of school you went to so I just ignored it. I chose those bounds because I think there is a 99 percent chance you can be accepted into a competitive chance but it has a high threshold and I think there is a chance you can get a 0 percent chance of getting into one too.

(c)

NOTE: THIS WAS BEFORE I REALIZED I HAD TO MODEL IT WITH THE REGRESSOR

ACT: The higher the ACT score, the higher the acceptance rate goes too. This makes sense because the ACT is there to gauge your skills to get into high school so the better you do the more chance that you can get in.

Extracurricular: This score is either 0 or 1, with 0 being a negative impact and 1 being a positive impact. You can get in without one, the score being 0, but it will make a good acceptance rate a lot harder to achieve. This makes sense because colleges like to see you being active and not solely relying on academics, you can bring more to the table than just your grades.

Interaction Term(ACT * Extracurricular): As this variable increases, so do your chances. This interaction is essentially a cheat code because if you do an extra curricular it gives you more of a boost for your ACT score. You don't have to get the highest score but you still need a relatively high score. This makes sense because colleges want to see pro activeness and if they see it in an extracurricular they will be more lenient on the ACT score.

NOTE: This was my real numbers based on the model :'ACT': -2.46759874, 'extra': -31.58896066, 'ACT$_{e}xtra'$: 0.80404903,$'constant'$: 87.51397881

They're really weird and I don't know why **(d)**

['Martha', '33', '1', 'Accept'] = 0.999
['Jeremy', '34', '0', 'Reject']
['Alphie', '35', '1', 'Accept']
['Dennis', '30', '1', 'Reject']
['Jennifer', '36', '1', 'Accept']
['Martin', '29', '1', 'Reject']
['Mary', '36', '1', 'Accept']
['Dean', '31', '1', 'Reject']
['Adam', '36', '0', 'Accept']
['Jeremy', '32', '0', 'Reject']

Problem 3

Regression is when you measure the trend of certain data points, and fit a function to those trends.It measures relationship between the data to set a pattern between all of it. This can be used to establish relationships between known variables and uncertainties, like the probability of a heart attack occurring via known traits and a set of people who have had heart attacks. It can be used to possibly predict the future of an environment or problem by using what we already know. Regression is important because it allows us to predict the future, find relationships and explain why certain things occur or help us how to

go forward with our newfound data.

There are two types of regression I'm going to talk about, linear and logistic. Linear regression comes in the form of a straight line.

[insert line with data points around it]

Linear regression can be modeled with this equation

$$y = \beta_0 + \beta_1 x$$

Logistic is a form of regression that comes in a sigmoid shape and has an upper and lower limit. The reason it takes a sigmoid shape is because at the beginning it starts at the lower limit, but once it increases and goes towards the higher limit it levels out again, hence forming the s-like shape of the graph.

[insert sigmoid photo and data points around it]

It can be modeled with this type of equation

$$y = \frac{1}{1 + e^{\beta x}}.$$

This is the standard one-variable equations but both can be changed to fit multiple variables. Linear becomes:

$$y = \beta_0 + \beta_1 x_0 + \beta_2 x_1 + \beta_3 x_2 + ...$$

and logistic becomes

$$y = \frac{1}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_1 + ...}}.$$

Both regressions have their differences and fit in different scenarios. Logistic regression has a key difference to linear in which it has limits. These limits can mean a variety of different things, but they can help explain the impact of a relationship we would analyze. Say population of a species in a new environment. This would be a clear cut case of using logistic over linear because you can't have negative population so it would have a lower limit and you can't have the population of a species grow forever because of environmental limits. This is where we can choose logistic regression over linear regression. Linear regression as said before doesn't have limits and is used to model different scenarios, ones where a relationship is continuous, say you're modeling experience with something and how much time it will take. The more you do something and the more you get comfortable the faster and faster you get at it, or maybe the yield of crops depending on how much water or fertilizer you use. These relationships are much more simple and use less variables. They also don't really rely or need upper or lower limits.

Logistic regression is also really good for determining probabilities. The reason you can use it to measure probabilities comes in part due to their limits. These limits can be the probability of something happening and outside circumstances affect your probability. A specific example of this is weather. You can determine the chance of rain or sunshine using weather patterns of past months

and days to predict the likelihood of these things happening. Linear regression can't model probabilities because there are no limits. There isn't a 200% of something happening or a negative percent chance, also probability isn't always linear, it changes.

Logistic regression also doesn't have to be in the bounds of 0 to 1, though that is normal. Remember this equation:

$$y = \frac{1}{1 + e^{\beta x}}.$$

If you change the numerator, then that is how you can change the upper limit of your regression. So this equation presents itself, to change the upper limit:

$$y = \frac{K}{1 + e^{\beta x}}.$$

In this equation K is your upper limit. This can be used for say a movie rating system that goes from 0 to 10 or 0 to 5 stars. Another example of needing different limits can be determining the perfect amount of something. Like say amount of toppings on a pizza. You can relate the amount of toppings with customer satisfaction and determine an average amount of toppings that would lead to best reviews from customers. These are cases in which you can go from 0 to a number of different upwards bounds.

Not only can you change the upper bound but the lower bound is slightly different. Here you have to adjust equation again. Here is a generalized formula to fit a logistic regression for bounds of your choice.

$$y = k + \frac{K - k}{1 + e^{\beta x}}.$$

In this equation, K is your upper limit, and k is your lower limit. We would want to change the limits for different scenarios just like normal. Say you have two people going in certain directions. You want to find a patter of which directions people go left or right, or split. This could go from -1 to 1 where -1 represents left and right and the inputs would be the peoples interests and what are on the other sides. Here we can predict what direction people would choose depending on their interests using our regression. And remember this is based off of 2 people, which is why we have different bounds. This is just a one off scenario but there are many other scenarios too. Resolving the behavior of a stock in the stock market based on outside influences. Stocks range from negative to high so you can have your bounds be from -5 to 5. There are a ton of different scenarios in which you wouldn't want the standard 0 to 1 bounds.