CS188–Winter 2020 — Homework 2 Solutions

Hermmy Wang, SID 704978214

Collaborators: None

2. Perceptron Training

b	w1	w2	w3	x1	x2	x3	t	z	Accuracy	delta_b	delta_w1	delta_w2	delta_w3
1	1	1	1	1	0	1	0	1	50%	-1	-1	0	-1
				1	1	0	0	1					
				1	0	1	1	1					
				0	1	1	1	1					
0	0	1	0	1	0	1	0	0	50%	-1	-1	-1	0
				1	1	0	0	1					
				1	0	1	1	0					
				0	1	1	1	1					
-1	-1	0	0	1	0	1	0	0	50%	1	1	0	1
				1	1	0	0	0					
				1	0	1	1	0					
				0	1	1	1	0					
0	0	0	1	1	0	1	0	1	75%	-1	-1	0	-1
				1	1	0	0	0					
				1	0	1	1	1					
				0	1	1	1	1					
-1	-1	0	0	1	0	1	0	0	50%	1	1	0	1
				1	1	0	0	0					
				1	0	1	1	0					
				0	1	1	1	0					
0	0	0	2	1	0	1	0	1	75%	-1	-1	0	-1
				1	1	0	0	0					
				1	0	1	1	1					
				0	1	1	1	1					
-1	-1	0	1	1	0	1	0	0	50%	1	1	0	1
				1	1	0	0	0					
				1	0	1	1	0					
				0	1	1	1	0					
0	0	0	2	1	0	1	0	0	50%	1	1	0	1
				1	1	0	0	0					
				1	0	1	1	0					
				0	1	1	1	0					

3. Input Validation

We can use checksum to verify the integrity of data. Checksum is a calculated value that serves as a unique identifier for the input stream. If the data changes then so does the checksum value. This makes it easy to verify the integrity of the data. If the checksum is outside a specified range, it implies the input stream either has missing data fields or the input value is an outlier.
Another method is to explicitly specify a range for each data field and check the range for input data stream. If a data field is null or outside the range, it is easily detectable. The range for each data field can be found by investigating historical dataset which is already processed and guaranteed to falling into a reasonable range.

4. Distributions

(a) (a) 2 * (14% + 6%) = 40%

(b) Unknown because the interval of the distribution only has the value 66-68 inch. We have no idea of what percent of individuals is between 66-67 inch.

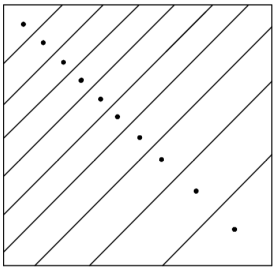
(c) The number of sons is unknown because we do not know the total number of sons.

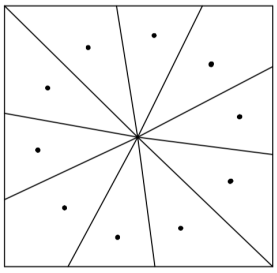
(d) 200 * (100% - 2*2%) = 200 * 96% = 192.

(b) The new bin contains 2*(2+8)=16% in a width of 4 inch. The height is 4 percent per inch.

(c) The tallest mother is between 70 and 72 inches. 2*(8+2)=20% of sons are above 72 inches. 2*(14+8+2)=48% of sons are above 70 inches. The percentage of sons that are taller than all of the mothers is between 20 and 48.

5. Voronoi





In these two diagrams, we have the following aspects in common. If we consider the entire plane boundless, the area of each region is infinite. In other words, each region extends forever. In addition, each region is bounded only by two lines.

6. Augmentation

(a) We can cross cholserum: Cholestoral in mg/dl with exang: Exercise induced angina. Cholestoral level in human body is related to many factors such as age and activity level of a person. A person who exercises a lot may have lower cholesterol level but in the meanwhile has a higher chance to get exercise induced angina (patients who never exercise have angina for some other reasons). A person who has a combination of high level of cholestoral and no exercise induced angina might have a higher impact than each feature on their own.
We can also cross fasting blood sugar with age. It is normal for elderly people who are over a certain age to have higher blood sugar level than younger people. This is because as we get old, blood sugar processing slows down and we do not use insulin as effectively as we do in younger ages. High blood sugar can damage blood vessels and cause heart disease. A combination of high fasting blood sugar and senior age may be worth more each feature on its own.

(b) Since we want to do feature crossing on home's number of bedrooms and location, we can divide each range of latitudes and longitudes into intervals/bins. If a home falls into a specific interval of latitude and longitude, we can assign a label of location for this home. Crossing binned latitude with binned longitude enables the model to the effect of location on the number of rooms. If a home falls into a specific interval of latitude and longitude, we can assign a label of location for this home.

(c) Suppose we have a linear model which is not linearly separable. The weights is set to b=0, w1=1, w2=2 such that $Z_1 = 1$ when $X + Y > 0$ and $Z_1 = -1$ otherwise. The linear model performs poorly (Accuracy=62.5%) because the model is not linearly separable. When we do feature crossing XY and let $Z_2 = 1$ when $XY > 0$ and $Z_2 = -1$ otherwise, the model performs well (Accuracy=100%).

X	Y	X+Y	Z_1	XY	Z_2
5	-1	4	1	-5	-1
-2	1	-1	-1	-2	-1
4	-3	1	1	-12	-1
-4	1	-3	-1	-4	-1
1	2	3	1	2	1
-2	-2	-4	-1	4	1
2	3	5	1	6	1
1	5	6	1	5	1