Module 12 Self Check

David Bishop

November 2024

1 Euclidean Distance

1. 3 Nearest Neighbors Calculate Distances for each point:
   1: $d(p_6, p_1) = \sqrt{(0.39 - 0.23)^2 + (0.63 - 0.81)^2} = 0.241$
   2: $d(p_6, p_2) = \sqrt{(0.39 - 0.42)^2 + (0.63 - 0.78)^2} = 0.153$
   3: $d(p_6, p_3) = \sqrt{(0.39 - 0.64)^2 + (0.63 - 0.23)^2} = 0.472$
   4: $d(p_6, p_4) = \sqrt{(0.39 - 0.87)^2 + (0.63 - 0.19)^2} = 0.651$
   5: $d(p_6, p_5) = \sqrt{(0.39 - 0.76)^2 + (0.63 - 0.43)^2} = 0.421$

   As we can see above, the closest points are 2 at 0.153, 1 at 0.241, and 5 at 0.421.

2. Predicted y value for 6

   $y_{pred} = \frac{0.33+0.18+0.32}{3} = 0.277$

3. Necessity of Square Root

   You do not need the square root to compare distances. The square root is only to get the actual distance of each point. If you only need to compare them then the un-square rooted distance will suffice as the relative distance between the points should remain the same.

2 Performance Metrics

1. accuracy

   $\frac{TP+TN}{TP+TN+FP+FN} = \frac{329+357}{329+357+35+87} = 0.8495$

2. error

   $\frac{FP+FN}{TP+TN+FP+FN} = \frac{35+87}{329+357+35+87} = 0.1505$

3. precision

   $\frac{TP}{TP+FP} = \frac{329}{329+35} = 0.9044$

4. recall

   $\frac{TP}{TP+FN} = \frac{329}{329+87} = 0.7909$

3 Mean Squared Error

y	y_{hat}	$y - y_{hat}$	$(y - y_{hat})^2$
3.78	3.32	0.46	0.2116
4.82	5.21	-0.39	0.1521
2.83	2.97	-0.14	0.0196
2.76	2.37	0.39	0.1521
3.48	3.29	0.19	0.0361

1. MSE
   $\frac{0.2116+0.1521+0.0196+0.1521+0.0361=0.5715}{5} = 0.1143$

2. Mean of y

   $\frac{3.78+4.82+2.83+2.76+3.48}{5} = 3.534$

3. MSE using Mean of y

   $(3.78 - 3.534)^2 = 0.0601$
   $(4.82 - 3.534)^2 = 1.6496$
   $(2.83 - 3.534)^2 = 0.495$
   $(2.76 - 3.534)^2 = 0.5971$
   $(3.48 - 3.534)^2 = 0.0029$
   $0.0601 + 1.6496 + 0.4950 + 0.5971 + 0.0029 = 2.8047$
   $\frac{2.8047}{5} = 0.5609$

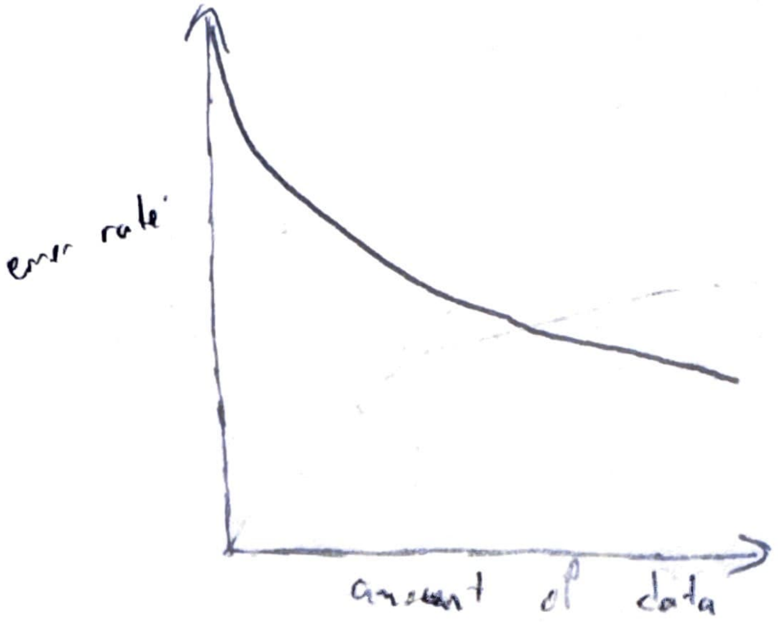
4. Variance of population

   Based on the question I am unsuprised, but overall it is definitely interesting to see what small tweaks in a formula can cause them to be able to calculate something else

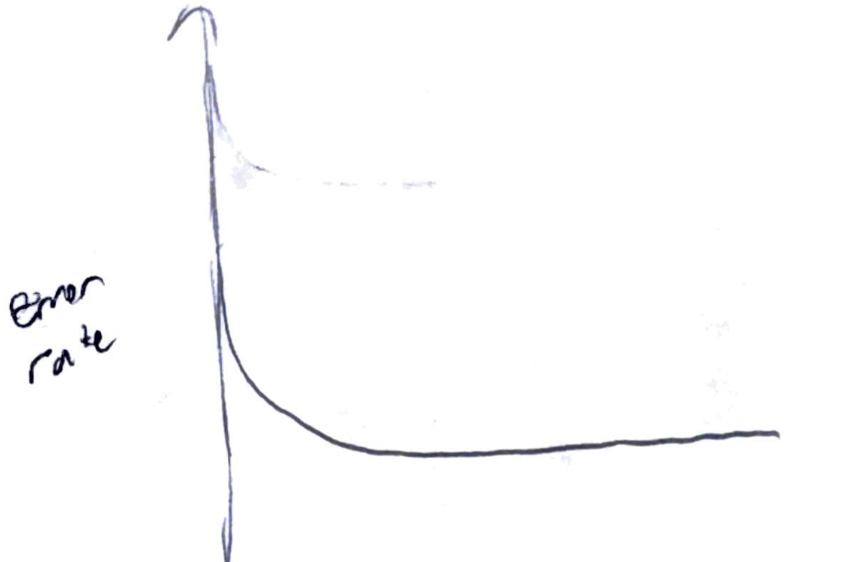
4 Learning Curves

Next page, The first one should be a curve where the error rate would still be going down as the data is added. For the second it should be that the curve has flattened where adding more data is no longer decreasing the error rate.

1. More data may help



2. More data won't help


error rate
amount of data