Virpobe Paireepinart

Homework 2 for Advanced Data Mining

Spring 2009 Texas State University

1) Naive Bayesian Classification of "Systems", "26...30", "46k...50k"

Total number of training data: 30 + 40 + 40 + 20 + 5 + 3 + 3 + 10 + 4 + 4 + 6 = 165

---

Per-class ("status") probabilities:

P(status = "senior") = 30 + 5 + 3 + 10 + 4 / 165 = 52 / 165 = 0.3152

P(status = "junior") = 40 + 40 + 20 + 3 + 4 + 6 / 165 = 113/165 = 0.685

---

Per-value ("department", "age", "salary") probabilities, for input

X = (department = "systems", age = "26....30", salary = "46K...50K")

(Using Laplacian Correction)

P(department = "systems" | status = "senior") = 5 + 3 + 1 / 52 = 9 / 52 = 0.173

P(age = "26...30" | status = "senior") = 0 + 1 / 52 = 1 / 52 = 0.0192

P(salary = "46K...50k" | status = "senior") = 30 + 10 + 1 / 52 = 41 / 52 = 0.788

P(department = "systems" | status = "junior") = 20 + 3 + 1 / 113 = 24 / 113 = 0.212

P(age = "26...30" | status = "junior") = 40 + 3 + 6 + 1 / 113 = 50 / 113 = 0.442

P(salary = "46K...50K" | status = "junior") = 20 + 3 + 1 / 113 = 24 / 113 = 0.212

---

P(X|status = "senior") = 0.173 \* 0.0192 \* 0.788 = 0.003

P(X|status = "junior") = 0.212 \* 0.442 \* 0.212 = 0.020

---

P(X|status = "senior") \* P(status="senior") = 0.003 \* 0.3152 = 0.001

P(X|status = "junior") \* P(status="junior") = 0.020 \* 0.685 = 0.014

---

Therefore, X is classified as a "junior"!

(note that you can intuit this pretty easily,

since no seniors are between the ages of 26 and 30 in the sample space,

so it's very unlikely that our input is a "senior.")

2.

We are given the data

18, 22, 25, 42, 28, 43, 33, 35, 56, 28;

(a)

The mean of this data set is:

18 + 22 + 25 + 42 + 28 + 43 + 33 + 35 + 58 + 28 / 10 = 332 / 10 = 33.2

The mean absolute deviation is the mean of the deviations of each value from the mean.

Therefore we can take each value, find out its (absolute) distance from the mean, add up all these distances,

and then take the mean of them.

33.2 - 18 = 15.2

33.2 - 22 = 11.2

33.2 - 25 = 8.2

33.2 - 42 = 8.8

33.2 - 28 = 5.2

33.2 - 43 = 9.8

33.2 - 33 = 0.2

33.2 - 35 = 1.8

33.2 - 56 = 22.8

33.2 - 28 = 5.2

Then we sum these up...

15.2 + 11.2 + 8.2 + 8.8 + 5.2 + 9.8 + 0.2 + 1.8 + 22.8 + 5.2 = 88.4

And we divide by the total number of items to get the final mean:

88.4 / 10 = 8.84

So the mean absolute deviation for this data set is: 8.84

(b) Z-score for the first four measurements:

Z score is simply (X - mean) / mean absolute deviation.

So for the first four measurements the Z-score is:

(18 - 33.2) / 8.84 = -1.719

(22 - 33.2) / 8.84 = -1.267

(25 - 33.2) / 8.84 = -0.928

(42 - 33.2) / 8.84 = 0.995

3.

Given two objects (22, 1, 42, 10) and (20, 0, 36, 8),

(a) compute euclidean distance:

euclidean distance is regular distance. just get difference

between corresponding points, square the differences,

add them together and take square root.

therefore:

d(p, q) = sqrt((22 - 20)\*\*2 + (1 - 0)\*\*2 + (42 - 36)\*\*2 + (10 - 8)\*\*2)

= sqrt( 2\*\*2 + 1\*\*2 + 6\*\*2 + 2\*\*2) = sqrt( 4 + 1 + 36 + 4) = sqrt(45) = 6.7082

(b) compute the manhattan distance.

manhattan distance is just the sum of the deltas of each coordinate set.

d(p, q) = abs(22-20) + abs(1 - 0) + abs(42 - 36) + abs(10 - 8) = 2 + 1 + 6 + 2 = 11

(c) compute the minkowski distance, q=3.

You can think of minkowski distance as generalization of previous two distance metrics.

basically it's p\_root(sum(abs(x - y)\*\*p) for x,y in P, Q)

d(p,q) = cuberoot(abs(22-20)\*\*3 + abs(1-0)\*\*3 + abs(42-36)\*\*3 + abs(10-8)\*\*3)

= cuberoot(2\*\*3 + 1\*\*3 + 6\*\*3 + 2\*\*3)

= cuberoot(8 + 1 + 216 + 8)

= cuberoot(233)

= 6.1534