**Problem set 2 solutions**

**Problem 1**

Notes:

Likelihood = P(Feature | Class)

Evidence = P(Feature) = sum(i = 1,..., n) [ P(Feature | Classi) \* P(Classi) ]

Prior probability = P(Class)

Posterior probability = P(Class | Feature)

Baye’s Theorem:

P(Class | Feature) = P(Feature | Class) \* P(Class) / P(Feature)

(a) *Suppose we want to classify sequence fragments into categories (represented by random variable Y): genes, regulatory motifs, or repetitive elements. We want to use the following features: length X1 , GC content (proportion of bases which are G or C) X2 , and complexity X3 (intuitively, what fraction of possible k-mers are observed). Does the naive Bayes assumption hold in this setting? Explain why or why not.*

The assumption is that for each class of Y, all variables (Xi) are independent. To determine independence, we must think about how each X is formulated

X1 = total number of bases in sequence = Ntot

X2 = (number Gs + number Cs) / total length = (Nc + Ng) / Ntot = (Nc + Ng) / X1

X3 = (number hits of given k-mer) / (total number of k-mers in sequence)

= k / (Ntot - k)

= k / (X1 - k)

Since X2 and X3 depend on X1, the Naive Bayes assumption does not hold true.

(b) Calculate conditional probabilities from the table:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Xi | Value of Xi (Vj) | Gene | Motif | Repeat | P(X1 = Vj | Gene) | P(X2 = Vj | Motif) | P(X3 = Vj | Repeat) |
| GC content | Low | 2 | 0 | 0 | 2/3 | 0/4 | 0/3 |
| Med | 1 | 2 | 0 | 1/3 | 2/4 | 0/3 |
| High | 0 | 2 | 3 | 0/3 | 2/4 | 3/3 |
| Length | Short | 0 | 4 | 0 | 0/3 | 4/4 | 0/3 |
| Long | 3 | 0 | 3 | 3/3 | 0/4 | 3/3 |
| Complexity | Low | 1 | 1 | 2 | 1/3 | 1/4 | 2/3 |
| High | 2 | 3 | 1 | 2/3 | 3/4 | 1/3 |

We can then calculate the conditional probability for a given set of conditions with the formula:

We can also calculate the prior probabilities:

P(Gene) = 3/10 = 0.3

P(Motif) = 4/10 = 0.4

P(Repeat) = 3/10 = 0.3

(c) To calculate the posterior estimate from the observations

X1 = Medium

X2 = Long

X3 = Low

We can use the Maximum-A-Posteriori estimates

Best class = maxY( P(Y | Xi) )

= maxY( P(Xi| Y) \* P(Y) )

Given that

P(Xi| Y) = P(X1 | Y) \* P(X2 | Y) \* ... \* P(Xn | Y)

= product(i = 1,..., n) [ P(Xi | Y) ]

P(Gene | X1 = MED, X2 = LONG, X3 = LOW)

= P(X1 = MED | Gene) \* P(X2 = LONG| Gene) \* P(X3 = LOW | Gene)

= (1/3) \* (3/3) \* (1/3)

= 3/27

= 0.111

P(Motif | X1 = MED, X2 = LONG, X3 = LOW)

= P(X1 = MED | Motif) \* P(X2 = LONG| Motif) \* P(X3 = LOW | Motif)

= (2/4) \* (0/4) \* (1/4)

= 0/27

= 0

P(Repeat | X1 = MED, X2 = LONG, X3 = LOW)

= P(X1 = MED | Repeat) \* P(X2 = LONG| Repeat) \* P(X3 = LOW | Repeat)

= (0/3) \* (3/3) \* (2/3)

= 0/27

= 0

Best class = maxY( 0.111\*0.3, // gene

0\*0.4, // motif

0\*0.3 // repeat

)

= 0.111, Gene

**Problem 2**

Conditional probability distribution for two states, conserved (C) and non-conserved (N)

|  |  |  |
| --- | --- | --- |
| Score | P(S|N) | P(S|C) |
| 0 | 0.1 | 0.05 |
| 1 | 0.35 | 0.15 |
| 2 | 0.25 | 0.2 |
| 3 | 0.2 | 0.3 |
| 6 | 0.1 | 0.3 |

1. Compute the conditional probability of each of the following alignments given each of the models

ACGACGACTA

CAGACGCTGA

TTCCTCTGAT

AGATGTGACT

-------------

2022220002

P(Si|N) = 0.25\*0.1\*0.25\*0.25\*0.25\*0.25\*0.1\*0.1\*0.1\*0.25

= 0.256 \* 0.14

=2.44e-8

P(Si|C) = 0.2\*0.05\*0.2\*0.2\*0.2\*0.2\*0.05\*0.05\*0.05\*0.2

= 0.26 \* 0.054

= 2.44e-8

= 4e-10

Since P(Si|N) > P(Si|C), the alignment is non-conserved.

ACAACGAGTA

AAAACGAATA

TCATCGAGTT

ACATCTAACT

----------------

3362636232

P(Si|N) = 0.253 \* 0.24 \* 0.13

= 2.5e-8

P(Si|C) = 0.23 \* 0.34 \* 0.33

= 1.75e-6

Since P(Si|C) > P(Si|N), the alignment is conserved.

(b) Generating samples from N, P(S|C) > P(S|N) 12.86% of the time

(c) Generating samples from C, P(S|N) > P(S|C) 14.16% of the time

(d) To reduce the rate of classification errors, we could increase N, or change the probability distributions of N and C, such that N skews more towards less pairs and C skews more towards more pairs.

**Problem 3**

(a) See code

(b) See plots

(c) See plots.

*What when wrong?*

There is a group of points in the lower right quadrant that got grouped with the green cluster. This group of points looks like it could be it’s own cluster (ie. K should be increased to 4).

*What strategy would I use to dynamically set the parameters so that K is chosen correctly?*

Solution 1: Choose K dynamically

<https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb>

1. The elbow method - calculate the Within-Cluster Sum of Squared Errors (WSS) for a variety of K values. Choose the K value at the elbow of the WSS vs K graph.
2. The silhouette method - calculate the Silhouette Score for each point for a variety of K values. The Silhouette Score (SS) is a measure the similarity and dissimilarity of that point to other points in the cluster. A higher SS indicates points are in the right clusters. I assume take the average SS for all points and graph SS vs. K. The optimal K value will be the peak SS.

Solution 2: Initialize centers based on data, not randomly

Solution 3: Set the stopCriteria distance to be smaller

Based on the rest of the question (ie. You only need to change one line), it looks like either 2 or 3 is the solution they are thinking of.

Attempts

-Tried solution 3: Didn’t work

-Tried solution 2: Initialized centers as random points in the data table -> this worked. In fact for this distribution K = 3 seemed appropriate.