Assignment 2

CS 534: Machine Learning

1. In DNA, also known as the Code of Life, there exist four different possible bases: adenine (abbreviated A), cytosine (C), guanine (G) and thymine (T). We are given an organism of unknown DNA base frequencies. Let p_a; p_c; p_g, and p_t be those unknown frequencies. Assume that we have obtained a strand of DNS sequences and we want to estimate the unknown frequencies. Let n_a; n_c; n_g; n_t be the corresponding number of bases that you observe for A, C, T and G respectively. Please derive the maximum likelihood estimates for the unknown parameters p_a; p_c; p_g, and p_t.

   Answer:

   For simplification we use 1, 2, 3, 4 respectively for A, C, G and T.
   The probability of unknown frequencies is p_i with constraint $\sum_{i=1}^{4} p_i = 1$
   Total number of bases is $n, where\ n = \sum_{i=1}^{4} n_i$
   In this case the joint probability is a vector n_1, n_2, n_3, n_4 and is called multinomial and has the following form

$$p(n_1, n_2, n_3, n_4 \mid p_1, p_2, p_3, p_4) = \frac{n!}{\prod n_i!} \prod p_i^{n_i}$$

   Now we will find the log likelihood of the above function which is:

$$l\ (p_1, p_2, p_3, p_4) = \log n! - \sum_{i=1}^{4} \log n_i! + \sum_{i=1}^{4} n_i \log p_i$$

   Before maximizing this, we would like to impose the constraint and for that we need to use Lagrange multiplier again:

$$l\ (p_1, p_2, p_3, p_4, \lambda) = l\ (p_1, p_2, p_3, p_4) + \ \lambda(1 - \sum_{i=1}^{4} p_i)$$

   Now to find the maximum likelihood: we take partial derivative of l with respect to each parameter p_i and setting it to zero.
   $\frac{\delta\, l}{\delta_{p_i}} = \frac{n_i}{p_i} - \lambda$ for each i = 1, 2, 3, 4.
   To maximize this, we use the constraint $\sum_{i=1}^{4} p_i = 1$ and conclude $\lambda = n$

   $\frac{n_i}{p_i} - n = 0\ that\ leads\ to\ p_i = \frac{n_i}{n}$ for each i = 1, 2, 3, 4

$$(ans.)$$

2. Consider the following training set:

A	B	C	Y
0	1	1	0
1	1	1	0
0	0	0	0
1	1	0	1
0	1	0	1
1	0	1	1

(a) Learn a Naive Bayes classifier by estimating all necessary probabilities.

Answer:

From the training set we find the following probabilities:

Prior Probabilities: $P(Y = 1) = P(Y = 0) = \frac{1}{2}$

Conditional probability of classes given for Y = 1.

$P(A = 0|Y = 1) = \frac{1}{3}$ $P(A = 1|Y = 1) = \frac{2}{3}$

$P(B = 0|Y = 1) = \frac{1}{3}$ $P(B = 1|Y = 1) = \frac{2}{3}$

$P(C = 0|Y = 1) = \frac{2}{3}$ $P(C = 1|Y = 1) = \frac{1}{3}$

Conditional probability of classes given for Y = 0.

$P(A = 0|Y = 0) = \frac{2}{3}$ $P(A = 1|Y = 0) = \frac{1}{3}$

$P(B = 0|Y = 0) = \frac{1}{3}$ $P(B = 1|Y = 0) = \frac{2}{3}$

$P(C = 0|Y = 0) = \frac{1}{3}$ $P(C = 1|Y = 0) = \frac{2}{3}$

(b) Compute the probability of $P(Y = 1|A = 1, B = 0, C = 0)$.

Answer:

$$P(Y = 1|A = 1, B = 0, C = 0)$$

$$= \frac{P(A = 1, B = 0, C = 0|Y = 1)\, P(Y = 1)}{P(A = 1, B = 0, C = 0|Y)}$$

$$= \frac{P(A = 1, B = 0, C = 0|Y = 1)\, P(Y = 1)}{P(A = 1, B = 0, C = 0|Y = 1) + P(A = 1, B = 0, C = 0|Y = 0)}$$

$$= \frac{\frac{2}{3}*\frac{1}{3}*\frac{2}{3}*\frac{1}{2}}{\frac{2}{3}*\frac{1}{3}*\frac{2}{3}*\frac{1}{2} + \frac{1}{3}*\frac{1}{3}*\frac{1}{3}*\frac{1}{2}} = \frac{\frac{4}{54}}{\frac{4}{54} + \frac{1}{54}} = \frac{\frac{4}{54}}{\frac{5}{54}} = \frac{4}{5}$$ (ans.)

(c) Suppose we know that A, B and C are independent random variables, can we say that the Naïve Bayes assumption is valid?

Answer:

No. If A, B, C are independent then random variables then we can write the following:

P (A, B, C |Y) = P(A)*P(B)*P(C)

But in naïve bias we take advantage of the conditional probability assuming that A,B,C is independent of each other given Y, which results into the following:

P (A, B, C |Y) = P(A|Y) *P(B|Y) * P (C | Y)

3. As discussed in class, consider using a beta prior *Beta* (2*;* 2) for estimating *p*, the probability of head for a weighted coin. What is the posterior distribution of *p* after we observe 5-coin tosses and 2 of them are head? What is the posterior distribution of *p* after we observe 50-coin tosses and 20 of them are head? Plot the pdf function of these two posterior distributions. Assume that *p* = 0.4 is the true probability, as we observe more and more coin tosses from this coin, what do you expect to happen to the posterior?

Answer:

In this example we have a prior that is defined by beta (2, 2) distribution. So here

$$\alpha = 2 \; and \; \beta = 2$$

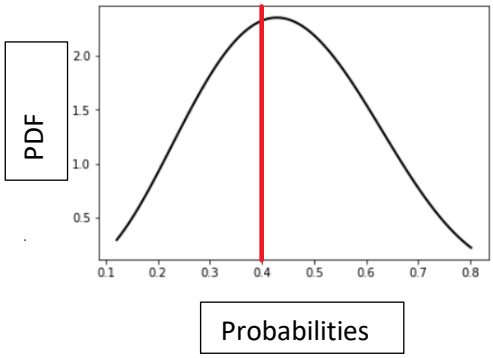
For 5-coin tosses with 2 heads our $n1 = 2 \; and \; n0 = 3$ where n0 is the number of times we get tails. The posterior distribution in this case is:

$$p(\theta|D) = \frac{1}{B(n1+\alpha, n0+\beta)}\,\theta^{n1+\alpha-1}\theta^{n0+\beta-1}$$

$$= \frac{1}{B(2+2,3+2)}\,\theta^{2+2-1}\theta^{3+2-1} = \frac{1}{B(4,5)}\,\theta^{3}\theta^{4}$$

Where, $\theta_{MAP} = \frac{n1+\alpha-1}{n+\alpha+\beta-2} = \frac{n1+1}{n+2} = \frac{2+1}{5+2} = \frac{3}{7} = .4286$

Here is the PDF function for the posterior generated above:



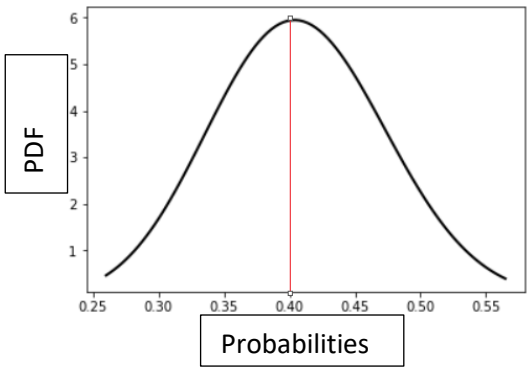
For 50-coin tosses with 20 heads our $n1 = 20\ and\ n0 = 30$ where n0 is the number of times we get tails. The posterior distribution in this case is:

$$p(\theta|D) = \frac{1}{B(n1+\alpha, n0+\beta)}\,\theta^{n1+\alpha-1}\theta^{n0+\beta-1}$$

$$= \frac{1}{B(2+20,30+2)}\,\theta^{2+20-1}\theta^{30+2-1} = \frac{1}{B(22,32)}\,\theta^{21}\theta^{31}$$

Where, $\theta_{MAP} = \frac{n1+\alpha-1}{n+\alpha+\beta-2} = \frac{n1+1}{n+2} = \frac{20+1}{50+2} = \frac{21}{52} = .4085$

Here is the PDF function for the posterior generated above:



From the figure above, we can see that as more tosses are happening we are reaching toward our true probability which is 0.4.

4. (Perceptron) The perceptron algorithm will only converge if the data is linearly separable. It is possible to *force* your data to be linearly separable as follows. If you have N data points in D dimensions, map data point $\vec{x}_n$ to the $(D+N)$-dimensional point $\langle \vec{x}_n, e_n \rangle$, where e_n is a N-dimensional vector of all zeros but one 1 at the nth position. (Eg., $e_4 = \langle 0, 0, 0, 1, 0, \ldots \rangle$.)

(a) Show that if you apply this mapping the data becomes linearly separable.

Answer:

If we apply this mapping, we will have our data set to be in a (D + N) dimension space. Now if we have a N data in N-d space each of the data is separable by a (N-1) dimension hyperplane. So, given we have (N+D) > N-1 dimensions we will be easily able to linearly separate the dataset. Now to achieve this we can construct a weight vector w, in (D + N) dimension space. We can initialize the weight vector as the first D points to be zero and then next n points to be +1 if it corresponds to a positive feature and -1 if it corresponds to a negative feature.

(b) How does this mapping affect generalization?

Answer:

This perceptron in (D+N) hyperplane will learn for every point in the dataset about the dimension it exists which doesn't give us any generalization capability because other features aren't involved.

5. In class, we showed that the quadratic kernel $K(\mathbf{x_i}; \mathbf{x_j}) = (\mathbf{x_i} \cdot \mathbf{x_j} + 1)^2$ was equivalent to mapping each $x = (\mathbf{x_1}; \mathbf{x_2}) \in \mathrm{R}^2$ into a higher dimensional space where

$$\Phi(\mathbf{x}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1).$$

Now consider the cubic kernel $K(\mathbf{x_i}; \mathbf{x_j}) = (\mathbf{x_i} \cdot \mathbf{x_j} + 1)^3$. What is the corresponding Φ function?

Answer:

Let, $(\mathbf{x_i}; \mathbf{x_j}) = \{(\boldsymbol{x_1}\,;\, \boldsymbol{x_2}), (\boldsymbol{x_1'}; \boldsymbol{x_2'})\}$

So, $K(\mathbf{x_i}; \mathbf{x_j}) = K((\boldsymbol{x_1}\,;\, \boldsymbol{x_2}).(\boldsymbol{x_1'}; \boldsymbol{x_2'}))^3 = K(1 + \boldsymbol{x_1}\,\boldsymbol{x_2} + \boldsymbol{x_1'}\boldsymbol{x_2'})^3$

$= (1 + \boldsymbol{x_1}\,\boldsymbol{x_1'} + \boldsymbol{x_2}\,\boldsymbol{x_2'})^3$

$= ((1 + \boldsymbol{x_1^2}\boldsymbol{x_1'^2} + 2\boldsymbol{x_1}\,\boldsymbol{x_1'}) + \boldsymbol{x_2}\,\boldsymbol{x_2'})^2(1 + \boldsymbol{x_1}\,\boldsymbol{x_1'} + \boldsymbol{x_2}\,\boldsymbol{x_2'})$

$= (\boldsymbol{x_1^2}\boldsymbol{x_1'^2} + 2\boldsymbol{x_1}\,\boldsymbol{x_2}\,\boldsymbol{x_1'}\boldsymbol{x_2'} + \boldsymbol{x_2^2}\boldsymbol{x_2'^2} + 2\boldsymbol{x_1}\,\boldsymbol{x_2} + +2\boldsymbol{x_1'}\boldsymbol{x_2'} + \mathbf{1})\ (1 + \boldsymbol{x_1}\,\boldsymbol{x_1'} + \boldsymbol{x_2}\,\boldsymbol{x_2'})$

$=(\boldsymbol{x_1^3}\boldsymbol{x_1'^3} + 3\,\boldsymbol{x_1^2}\boldsymbol{x_1'^2} + 3\boldsymbol{x_2^2}\boldsymbol{x_2'^2} + 3\boldsymbol{x_1^2}\boldsymbol{x_1'^2}\boldsymbol{x_2}\,\boldsymbol{x_2'} + 3\boldsymbol{x_1}\,\boldsymbol{x_1'}\boldsymbol{x_2^2}\boldsymbol{x_2'^2} + \mathbf{3}\boldsymbol{x_1}\,\boldsymbol{x_1'} + \mathbf{3}\boldsymbol{x_2}\,\boldsymbol{x_2'} + \mathbf{6}\boldsymbol{x_1}\,\boldsymbol{x_2}\,\boldsymbol{x_1'}\boldsymbol{x_2'} + \mathbf{3}\boldsymbol{x_2^3}\boldsymbol{x_2'^3} + \mathbf{1})$

$=(x_1^3 + \sqrt{3}x_1^2 + \sqrt{3}x_2^2 + \sqrt{3}x_1^2\ x_2 + \sqrt{3}x_1\ x_2^2 + \sqrt{3}x_1 + \sqrt{3}x_2 + \sqrt{6}x_1\ x_2 + \sqrt{3}x_2^3 + \mathbf{1}) \cdot (x_1'^3 + \sqrt{3}x_1'^2 + \sqrt{3}x_2'^2 + \sqrt{3}x_1'^2\ x_2' + \sqrt{3}x_1'\ x_2'^2 + \sqrt{3}x_1' + \sqrt{3}x_2' + \sqrt{6}x_1'x_2' + \sqrt{3}x_2'^3 + \mathbf{1})$

So,

$\Phi(x) = (x_1^3, \sqrt{3}x_1^2, \sqrt{3}x_2^2, \sqrt{3}x_1^2\ x_2, \sqrt{3}x_1\ x_2^2, \sqrt{3}x_1, \sqrt{3}x_2, \sqrt{6}x_1\ x_2, \sqrt{3}x_2^3, \mathbf{1})$

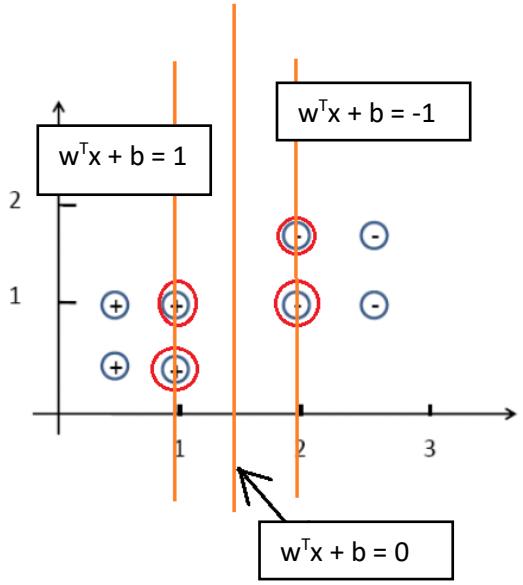
(ans.)

6. Apply linear SVM without soft margin to the following problem. Note that the two right most positive points are (1; 0:5) and (1; 1). The two left most negative points are (2; 1) and (2; 1:5).

(a) Please mark out the support vectors, the decision boundary ($\mathbf{w}^T\mathbf{x} + b = 0$) and $\mathbf{w}^T\mathbf{x} + b = 1$ and $\mathbf{w}^T\mathbf{x} + b = -1$. Note that you don't need to solve the optimization problem for this, just eyeball the solution.

Answer:

In the following image, the points with red circle around them are the support vectors which are the closest points to the boundary.



(c) Please solve for **w** and *b* based on the support vectors you identified in (a)

Answer:

We see point (1, 1) passing through $wx + b = 1$; so, $b = 1 - w$

We see point (2, 1) passing through $wx + b = -1$; so, $2w + 1 - w = -1$. Therefore $w = -2$.

If $w = -2$ plugging it for equation passing through (1, 1): $-2 + b = 1$, therefore $b = 3$.

$w = -2$ and $b = 3$.