Machine Learining Fall 2016

Sonam Choudhary

December 6, 2016

1 Probabilities

1. [2 points] **Solution 1 :** Yes A_1 and A_2 are independent events.This is concluded from the definition of independent events.Two events are said to be independent if the occurrence of one does not affect the probability of the occurence of the other.Mathematically this means if $P(A_1 \mid A_2) = P(A_1)$ and if $P(A_2 \mid A_1) = P(A_2)$ we say A_1 and A_2 are independent events.

2. [3 points] **Solution 2:** $P(A_4) = \sum_{i=1}^{3} P(A_4|A_i)P(A_i)$
$P(A_4) = (P(A_4|A_1)P(A_1))(P(A_4|A_2)P(A_2))(P(A_4|A_3)P(A_3))$
$P(A_4) = \frac{1}{6} * \frac{1}{3} * \frac{2}{6} * \frac{1}{3} * \frac{3}{6} * \frac{1}{3}$
$P(A_4) = \frac{6}{18} = \frac{1}{3}$

3. [3 points] **Solution 3:** Probability of getting a number(either of 1,2,3,4,5,6) on a die $= \frac{1}{6}$
We can denote this number by X.Here according to the question since we want atleast two heads so X can take values greater than or equal to 6.And let getting exactly two heads be event Y and the proability of getting a head or tail is $\frac{1}{2}$.We can use Binomial distribution as follows to calculate the probabilty of getting two heads given X:
$P(Y = 2|X = i) = C_2^i * \frac{1}{2}^i$ for i greater than equal to 2.
From conditional probability we know,
$P(Y = a) = \sum_{i=2}^{6} P(Y = a|X = i) * P(X = i)$,here a=2.
From above we can calculate $P(Y = 2) = \frac{1}{6} * \sum_{i=2}^{6} P(Y = 2|X = i) * P(X = i)$
$= 1/6(C_2^2 * 2^{-2} + C_2^3 * 2^{-3} + C_2^4 * 2^{-4} + C_2^5 * 2^{-5} + C_2^6 * 2^{-6})$
$= 1/6(1/4 + 3/8 + 6/16 + 10/32 + 15/64)$
$= 1/6(99/64)$
$P(gettingexactly2heads) = 33/128$

4. [4 points] **Solution 4:** we know Probability can take up values less than or equal to 1.So considering $P(A_1 \cup A_2)$
we can say $P(A_1 \cup A_2) \leq 1$
and $P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$

$\implies P(A_1) + P(A_2) - P(A_1 \cap A_2) \leq 1$
dividing both sides by $P(A_2)$ we get
$\implies \frac{P(A_1)+P(A_2)-P(A_1 \cap A_2)}{P(A_2)} \leq \frac{1}{P(A_2)}$
we know $P(A_1|A_2) = \frac{P(A_1 \cap A_2)}{P(A_2)}$
using the above two equations we get
$\frac{P(A_1)+P(A_2)}{P(A_2)} - \frac{P(A_1 \cap A_2)}{P(A_2)} \leq \frac{1}{P(A_2)}$
putting values we get:
$\frac{a_1+a_2}{a_2} - P(A_1|A_2) \leq \frac{1}{a_2}$
shifting terms we prove that
$P(A_1|A_2) \geq \frac{a_1+a_2-1}{a_2}$

5. [8 points] If A_1 and A_2 are independent events, then show that

   (a) $E[A_1 + A_2] = E[A_1] + E[A_2]$
   Soution 5.1 : We know for any random variable by the definition the expectation is equal to :
   $E[A_1] = \sum x * P(A_1 = x)$
   $E[A_2] = \sum y * P(A_2 = y)$
   Here x and y denotes the values the random variable can take.
   Let the value of random varibale be defined over space S.
   Now let $A_1 + A_2 = X$
   so by definition we can say the E[X] will be:
   $\implies E[X] = \sum_{i \epsilon S} X(i)P(i)$
   $\implies \sum_{i \epsilon S}(A_1(i) + A_2(i)) * P(i)$
   $\implies \sum_{i \epsilon S}(A_1(i) * P(i) + \sum_{i \epsilon S}(A_2(i) * P(i)$
   $\implies E[A_1] + E[A_2]$
   Hence proved
   $E[A_1 + A_2] = E[A_1] + E[A_2]$

   (b) $var[A_1 + A_2] = var[A_1] + var[A_2]$
   Solution 5.2: By definition of variance we know variance of a random variable is given by:
   $var[A] = E[A^2] - E[A]^2$
   $X = A_1 - E[A_1]$
   $Y = A_2 - E[A_2]$
   since A_1 and A_2 are independent events so E[X]=0 and E[Y] =0
   $\implies var[X] = var[A_1] and var[Y] = var[A_2] \implies va[A_1 + A_2] = var[X + Y]$ using above equations and definition of variance we get
   $\implies = E[(X + Y)^2] - (E(X + Y))^2$
   $\implies = E[X^2 + Y^2 + 2XY] - 0$
   $\implies = E[X^2] + 2E[XY] + E[Y^2]$
   $\implies = var[X] + 0 + var[Y]$
   from above
   $\implies = var[A_1] + var[A_2]$

hence proved.

Here $E[\cdot]$ and $var[\cdot]$ denote the mean and variance respectively.

2 Naive Bayes

1. [Part 1]

   (a) [2 points] **Solution 1:**The values of $\hat{P}(x_1 \mid y)$ and $\hat{P}(y)$ would be the same as of the true distribution if we have infinite data drawn from this distribution.It can be supported with an example of an unbiased coin toss where the probabilities of getting a head and that of tail coverges to $\frac{1}{2}$ when it tossed infinite coins as opposed to when it is say tossed for 4 or 5 times where the probabilty of either of them could be greated than the other.So the values are :
   when $x_1 = 1$

   $$\widehat{P}(x_1 = 1|y = -1) = 0.2$$

   $$\widehat{P}(x_1 = 1|y = 1) = 0.9$$

   when $x_1 = -1$

   $$\widehat{P}(x_1 = -1|y = 1) = 0.1$$

   $$\widehat{P}(x_1 = -1|y = -1) = 0.8$$

   and

   $$\widehat{P}(y = 1) = 0.9$$

   $$\widehat{P}(y = -1) = 0.1$$

   (b) [6 points] **Solution :**
   when x=-1:
   $\widehat{P}(x_1 = -1|y = -1) \times \widehat{P}(y = -1) = 0.8 \times 0.1 = 0.08$
   $\widehat{P}(x_1 = -1|y = 1) \times \widehat{P}(y = 1) = 0.1 \times 0.9 = 0.09$
   when x=1:
   $\widehat{P}(x_1 = 1, y = -1) = \widehat{P}(x_1 = 1|y = -1) \times \widehat{P}(y = -1) = 0.2 \times 0.1 = 0.02$
   $\widehat{P}(x_1 = 1, y = 1) = \widehat{P}(x_1 = -1|y = 1) \times \widehat{P}(y = 1) = 0.9 \times 0.9 = 0.81$

Input x_1	$\hat{P}(x_1, y = -1)$	$\hat{P}(x_1, y = 1)$	**Prediction:** $y' = arg\max_y \hat{P}(x_1, y)$
-1	0.08	0.09	$0.09 > 0.08$ so $y' = 1$
1	0.02	0.81,	$0.81 > 0.02$ so $y' = 1$

   (c) [3 points]
   Solution :when x_1 = -1 ,predicted label $y' = 1$ and error would be when true label $y = -1$ and similarly when $x_1 = 1$,predicted label $y' = 1$ and error would be when true label $y = -1$.So the probabilities would be:

$P(y=-1, x_1=-1) = P(x_1=-1|y=-1)(y=-1) = 0.8 \times 0.1 = 0.08$
$P(y=-1, x_1=1) = P(x_1=1|y=-1)(y=-1) = 0.2 \times 0.1 = 0.02$
So error of the classifier would be $P(y' \neq y) = 0.08 + 0.02 = 0.1$

2. [Part 2] Now, suppose we have a binary classification problem with two features x_1, x_2 both of which can be -1 or 1. However, the second feature x_2 is actually identical to the first feature x_1. And we have the same true probabilities $P(x_1 \mid y)$ and $P(y)$ as in Part 1 above.

   (a) [1 point]
   Solution: x_1 and x_2 are not conditionally independent.

   (b) [8 points]
   Solution:
   Probabilities are as below:
   when $x_1 = -1$ and $x_2 = -1$

$$P(x_1=-1, x_2=-1, y=-1) = 0.8 \times 0.8 \times 0.1 = 0.064$$
$$P(x_1=-1, x_2=-1, y=1) = 0.1 \times 0.1 \times 0.9 = 0.009$$

   we get,
   $P(x_1=-1, x_2=-1, y=-1) > P(x_1=-1, x_2=-1, y=1)$

   when $x_1 = -1$ and $x_2 = 1$

$$P(x_1=-1, x_2=1, y=-1) = 0.8 \times 0.2 \times 0.1 = 0.016$$
$$P(x_1=-1, x_2=1, y=1) = 0.1 \times 0.9 \times 0.9 = 0.081$$

   we get,
   $P(x_1=-1, x_2=1, y=-1) < P(x_1=-1, x_2=1, y=1)$

   when $x_1 = 1$ and $x_2 = -1$

$$P(x_1=1, x_2=-1, y=-1) = 0.2 \times 0.8 \times 0.1 = 0.016$$
$$P(x_1=1, x_2=-1, y=1) = 0.9 \times 0.1 \times 0.9 = 0.081$$

   we get,
   $P(x_1=1, x_2=-1, y=-1) < P(x_1=1, x_2=-1, y=1)$

   when $x_1 = 1$ and $x_2 = 1$

$$P(x_1=1, x_2=1, y=-1) = 0.2 \times 0.2 \times 0.1 = 0.004$$
$$P(x_1=1, x_2=1, y=1) = 0.9 \times 0.9 \times 0.9 = 0.729$$

we get,
$P(x_1 = 1, x_2 = 1, y = -1) < P(x_1 = 1, x_2 = 1, y = 1)$

x_1	x_2	$\hat{P}(x_1, x_2, y = -1)$	$\hat{P}(x_1, x_2, y = 1)$	**Prediction:** $y' = arg\max_y \hat{P}(x_1, x_2, y)$
-1	-1	0.064	0.009	-1
-1	1	0.016	0.081	1
1	-1	0.016	0.081	1
1	1	0.004	0.729	1

(c) [3 points] **Solution:**
For this question no assumption is made regarding the conditional independency.Since x_1 and x_2 are same ,the value of $P(x_1|x_2, y)$ will be 1 always.When $x_1 = -1$ and $x_2 = -1$ and the predicted label is $y' = -1$.we say error occurs when true label is $y = 1$. So error is $P(x_1 = -1, x_2 = -1, y = 1) = 0.09 \times 0.1 = 0.09$.

Similarly when $x_1 = 1, x_2 = 1$ and predicted label is $y' = 1$ while true label $y = -1$.

Error in this case wil be: $P(x_1 = 1, x_2 = 1, y = -1) = 0.2 \times 0.1 = 0.02$
Hence we can say that

$$P(y' \neq y) = 0.09 + 0.02 = 0.11$$

(d) [2 points] **Solution :**Logistic regression classifier performs better than the Naive Bayes as there in no requirement on the features to be conditionally independent unlike Naive Bayes which makes an assumption that the features are conditionally independent and thus makes more number of mistakes.

3 [25 points, Extra Credit for the holidays] Naïve Bayes and Linear Classifiers

Solution :
considering our input be d dimensional vectors, $\mathbf{x} = (x_1, x_2, \ldots, x_d)^T$.We know the condition when Our classifier predicts label 1 is if:
$\frac{P(\mathbf{x}|y=1)P(y=1)}{P(\mathbf{x}|y=0)P(y=0)} \geq 1 - - - - - > Eq1$
According to Naive Bayes assumption, we know that

$$P(\mathbf{x}|y) = \prod_{j=0}^{d} P(x_j|y)$$

. Using this we can rewrite equation 1 as
$\frac{p}{1-p} \prod_{j=0}^{d} \frac{P(x_j|y=1)}{P(x_j|y=0)} \geq 1 - - - - - - > Eq2$

Here, $p = P(y = 1)$ and $1 - p = P(y = 0)$
From Gaussian Distribution we can write:
$P(x_j|y_i) = g(x_j, \mu_{y_i}, \sigma_{y_i}) = \frac{1}{\sigma_{y_i}\sqrt{2\pi}} e^{-\frac{(x_j - \mu_{y_i})^2}{2\sigma_{y_i}^2}} \quad ---------- > Eq3$

From equation 2 and equation 3 we get:

$$\frac{p}{1-p} \prod_{j=0}^{d} \frac{\frac{1}{\sigma_{y_1}\sqrt{2\pi}} e^{-\frac{(x_j - \mu_{y_1})^2}{2\sigma_{y_1}^2}}}{\frac{1}{\sigma_{y_0}\sqrt{2\pi}} e^{-\frac{(x_j - \mu_{y_0})^2}{2\sigma_{y_0}^2}}} \geq 1$$

.
It's given that σ is same for both the classes.So after cancelling constants we get:

$$\frac{p}{1-p} \prod_{j=0}^{d} e^{-\frac{(x_j - \mu_{y_1})^2}{2\sigma_{y_1}^2}} + e^{\frac{(x_j - \mu_{y_0})^2}{2\sigma_{y_0}^2}} \geq 1$$

.
Taking log of the above equation:

$$\log \frac{p}{1-p} + \sum_{j=0}^{d} \frac{(x_j - \mu_0)^2 - (x_j - \mu_1)^2}{2\sigma^2} \geq 0$$

Simplifying the above equation:

$$\log \frac{p}{1-p} + \frac{\mu_0^2 - \mu_0^2}{2\sigma^2} + \sum_{j=0}^{d} \frac{\mu_1 - \mu_0}{\sigma^2} x_j \geq 0$$

Considering $\log \frac{p}{1-p} + \frac{\mu_0^2 - \mu_0^2}{2\sigma^2}$ as b, and $\frac{\mu_1 - \mu_0}{\sigma^2}$ as w^T, we get:

$$b + \sum_{j=0}^{d} w^T x_j \geq 0$$

The above equation shows that Gaussian Naive Bayes is also a linear classifier.

4 Experiment

1. [5 points] **Solution :**
   Given $g(\mathbf{w}) = \log(1 + exp(-y_i \mathbf{w}^T \mathbf{x}_i))$.
   Considering an n-dimensional vector $\mathbf{w}$.Taking the derivative we get

$$\frac{\partial g}{\partial w_j} = \frac{\partial}{\partial w_j} \log(1 + exp(-y_i \mathbf{w}^T \mathbf{x}_i))$$

$$\frac{\partial g}{\partial w_j} = \frac{\partial}{\partial w_j} \log(1 + exp(-y_i \mathbf{w}^T \mathbf{x}_i))$$

$$\frac{\partial g}{\partial w_j} = \frac{1}{1 + exp(-y_i \mathbf{w}^T \mathbf{x}_i)} \frac{\partial}{\partial w_j}(1 + exp(-y_i \mathbf{w}^T \mathbf{x}_i))$$

$$\frac{\partial g}{\partial w_j} = \frac{1}{1 + exp(-y_i \mathbf{w}^T \mathbf{x}_i)} - y_i \mathbf{x}_i exp(-y_i \mathbf{w}^T \mathbf{x}_i)$$

Simplifying the equation we get:

$$\frac{\partial g}{\partial w_j} = \frac{-y_i \mathbf{x}_i}{1 + e^{y_i \mathbf{w}^T \mathbf{x}_i}}$$

2. [5 points]**Solution :**The objective where the entire dataset is composed of a single example, say (x_i, y_i) is

$$\underline{\mathrm{J}}(w) = \log(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)) + \frac{1}{\sigma^2} \mathbf{w}^T \mathbf{w}$$

Taking the partial derivative with respect to weight component w_j, we get :

$$\frac{\partial J(w)}{\partial w} = \frac{-y_i - y_i \mathbf{x}_{ij}}{(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)} + \frac{2}{\sigma^2} w_j$$

calculating for all 1 ... n, gradient is :
$\bigtriangledown J(w) = \frac{\partial J}{\partial w_1}, \frac{\partial J}{\partial w_2}, \ldots \frac{\partial J}{\partial w_n}$
$\implies = \frac{-y_i \mathbf{x}_i}{(1+\exp(-y_i \mathbf{w}^T \mathbf{x}_i)} + \frac{2}{\sigma^2} \mathbf{w}$
$\bigtriangledown J(w) = \frac{-y_i \mathbf{x}_i}{(1+\exp(-y_i \mathbf{w}^T \mathbf{x}_i)} + \frac{2}{\sigma^2} \mathbf{w}$

Solution : Pseudo code is as written below:
We know for stochastic gradient for each example we calculate the gradient and then make an update in the oppposite direction so as to reach the minima. For each example we update the weight vector.:
$\mathbf{w}_{t+1} = w_t - r_t \bigtriangledown J(\mathbf{w})$
and as calculated above :
$\bigtriangledown J(w) = \frac{-y_i \mathbf{x}_i}{(1+\exp(-y_i \mathbf{w}^T \mathbf{x}_i)} + \frac{2}{\sigma^2} \mathbf{w}$
so using the above two equation and reshuffling terma w get:
$\mathbf{w}_{t+1} = \mathbf{w}_t(1 - \frac{2r_t}{\sigma^2}) + \frac{r_t y_i}{(1+\exp(-y_i \mathbf{w}^T \mathbf{x}_i)} \mathbf{x}_i$

Here $r_t = \frac{r_0}{1 + \frac{r_0 t}{\sigma^2}}$
Using the above we can write the pseudo code as below:
Considering the data set to be a single example x_i, y_i and running the code for T epochs:
t is the t^{th} example

STep1 :**For epoch 1** *dots* **T**
Step2: **initialise weights, w =0 and t=1**
Step 3:**shuffle the data**

Step 4:**for each example drawn from dataset**
Step 5:**calculate** r_t
STep 6: $r_t = \frac{r_0}{1+\frac{r_0 t}{\sigma^2}}$
Step 7: **update the weight**,
$\mathbf{w}_{t+1} = \mathbf{w}_t(1 - \frac{2r_t}{\sigma^2}) + \frac{r_t y_i}{(1+\exp(-y_i \mathbf{w}^T \mathbf{x}_i)}\mathbf{x}_i$
Step 8: **t=t+1**
STep 9:**return w**

Steps are written for better clarification couldn't manage a better way usinf Latex.

3. [20 points] **Solution:**
Accuracy on the system is 84.602
with sigma =50
Epochs $[3, 5]$
Sigma Square $[10, 50, 100]$
I have assumed that cross validation results are not asked therefore restricting values to only a few here.However i have tried much more epochs and other values for sigma.

epoch = 3

sigma	*Accuracy*	
10	77.63	
50	84.602	
100	80.265	

epoch = 5

sigma	*Accuracy*	
10	77.63	
50	79.602	
100	78.003	

Below is the graph: