HOMEWORK 4

>>Abhay Kumar<<
>>9081403157<<

Instructions: Although this is a programming homework, you only need to hand in a pdf answer file. There is no need to submit the latex source or any code. You can choose any programming language, as long as you implement the algorithm from scratch.
Use this latex file as a template to develop your homework. Submit your homework on time as a single pdf file to Canvas. Please check Piazza for updates about the homework.

1 Best Prediction Under 0-1 Loss (14 pts)

Suppose the world generates a single observation $x \sim$ multinomial(θ), where the parameter vector $\theta = (\theta_1, \ldots, \theta_k)$ with $\theta_i \geq 0$ and $\sum_{i=1}^{k} \theta_i = 1$. Note $x \in \{1, \ldots, k\}$. You know θ and want to predict x. Call your prediction $\hat{x}$. What is your expected 0-1 loss:

$$\mathbb{E}\mathbb{1}[\hat{x} \neq x]$$

using the following two prediction strategies respectively? Prove your answer.
Strategy 1: $\hat{x} \in \arg\max_x \theta_x$, the outcome with the highest probability.

$\mathbb{E}\mathbb{1}[\hat{x} \neq x] = 1\mathbf{x}Prob(x \neq \hat{x}) + 0\mathbf{x}Prob(x = \hat{x})$
$= Prob(x \neq \hat{x})$
$= 1 - Prob(x = \hat{x})$
$= 1 - Prob(x \in \arg\max_x \theta_x)$

$$= 1 - \sum_{j \in \arg\max_x \theta_x} \theta_j$$

Strategy 2: You mimic the world by generating a prediction $\hat{x} \sim$ multinomial(θ). (Hint: your randomness and the world's randomness are independent)
$\mathbb{E}\mathbb{1}[\hat{x} \neq x] = 1\mathbf{x}Prob(x \neq \hat{x})$
$= 1 - Prob(x = \hat{x})$
for every possible x values, we need $\hat{x}$ to be same for the above probability, i.e. $\theta_x \theta_{\hat{x}} = \theta_x^2$ Now, we need to sum it over all possible values x can take.
$= 1 - \sum_{j=1}^{k} \theta_j^2$ (using independence of both $\hat{x}$ and x)

2 Best Prediction Under Different Misclassification Losses (12 pts)

Like in the previous question, the world generates a single observation $x \sim$ multinomial(θ). Let $c_{ij} \geq 0$ denote the loss you incur, if $x = i$ but you predict $\hat{x} = j$, for $i, j \in \{1, \ldots, k\}$. $c_{ii} = 0$ for all i. This is a way to generalize different costs on false positives vs false negatives from binary classification to multi-class classification. You want to minimize your expected loss:

$$\mathbb{E}c_{x\hat{x}}.$$

Derive your optimal prediction $\hat{x}$.

$\mathbb{E}c_{x\hat{x}} = \sum_{i=1}^{k} c_{i\hat{x}} Prob(x = i)$

$= \sum_{i=1}^{k} c_{i\hat{x}} \theta_i$

We need to minimize your expected loss to get the optimal prediction of $\hat{x}$, Suppose the expected loss is minimized at $\hat{x} = \hat{\jmath}$ then $\hat{j}$ can be represented as-

$\hat{\jmath} = \arg\min_{\hat{x}} \sum_{i=1}^{k} c_{i\hat{x}}\theta_i$

3 Language Identification with Naive Bayes (8 pts each)

Implement a character-based Naive Bayes classifier that classifies a document as English, Spanish, or Japanese - all written with the 26 lower case characters and space.
The dataset is languageID.tgz, unpack it. This dataset consists of 60 documents in English, Spanish and Japanese. The correct class label is the first character of the filename: $y \in \{e, j, s\}$.
We will be using a character-based multinomial Naïve Bayes model. You need to view each document as a bag of characters, including space. We have made sure that there are only 27 different types of printable characters (a to z, and space) – there may be additional control characters such as new-line, please ignore those. Your vocabulary will be these 27 character types. (Note: not word types!)

1. Use files 0.txt to 9.txt in each language as the training data. Estimate the prior probabilities $\hat{p}(y = e)$, $\hat{p}(y = j)$, $\hat{p}(y = s)$ using add-1 smoothing. Give the formula for add-1 smoothing in this case. Print the prior probabilities. (Hint: Store all probabilities here and below in $\log()$ internally to avoid underflow. This also means you need to do arithmetic in log-space. But answer questions with probability, not log probability.)

   prior probabilities
   $\hat{p}(y = e) = \frac{\text{count of documents with (y=e) + 1}}{\text{Total count of documents+Number of classes}}$

	probability
$\hat{p}(y = e) =$	$\frac{10+1}{30+3} = 0.3333333$
$\hat{p}(y = j) =$	$\frac{10+1}{30+3} = 0.3333333$
$\hat{p}(y = s) =$	$\frac{10+1}{30+3} = 0.3333333$

2. Using the same training data, estimate the class conditional probability (multinomial parameter) for English

   $$\theta_{c,e} := \hat{p}(c \mid y = e)$$

   for $c \in \{a, \ldots, z, space\}$. Again use add-1 smoothing. Give the formula for add-1 smoothing in this case. Print θ_e which is a vector with 27 elements.

   let character count vector of a training document, x_i be $m_i = [m_{i,1}, m_{i,2}, m_{i,3}, ..., m_{i,27}]$ and $\theta_e = [\theta_{1,e}, \theta_{2,e}, \theta_{3,e}, ..., \theta_{27,e}]$

   $\theta_{j,e} = \frac{\sum_{i=1}^{N_e} m_{i,j}+1}{\sum_{j=1}^{27}\sum_{i=1}^{N_e} m_{i,j}+27}$, where N_e is the total number of training examples with class label- "e"

   $\theta_{c,e} = \frac{\text{count of character c in all documents of class label e}+1}{\text{total count of all 27 characters in all documents of class label e}+27}$

$$\theta_e = \begin{bmatrix} 0.06014789 \\ 0.01115806 \\ 0.02152383 \\ 0.021986 \\ 0.10530833 \\ 0.0189489 \\ 0.01749637 \\ 0.04720718 \\ 0.05539416 \\ 0.00145253 \\ 0.00376337 \\ 0.02898455 \\ 0.02053347 \\ 0.05790308 \\ 0.06443946 \\ 0.0167701 \\ 0.00059422 \\ 0.05380959 \\ 0.06615608 \\ 0.08008715 \\ 0.02667371 \\ 0.00930939 \\ 0.01551565 \\ 0.00118843 \\ 0.01386505 \\ 0.00066024 \\ 0.1791232 \end{bmatrix}$$

3. Print θ_j, θ_s.

$$\theta_j = \begin{bmatrix} 0.13167632 \\ 0.01089157 \\ 0.0055156 \\ 0.01724499 \\ 0.06018292 \\ 0.0039098 \\ 0.01403337 \\ 0.03176709 \\ 0.09697689 \\ 0.0023738 \\ 0.05739021 \\ 0.00146617 \\ 0.03979613 \\ 0.05669203 \\ 0.0911122 \\ 0.00090763 \\ 0.00013964 \\ 0.0427983 \\ 0.04216994 \\ 0.0569713 \\ 0.07058577 \\ 0.00027927 \\ 0.01975843 \\ 0.00006982 \\ 0.01417301 \\ 0.00774977 \\ 0.12336801 \end{bmatrix}$$

$$\theta_s = \begin{bmatrix} 0.10450428 \\ 0.00825682 \\ 0.03752542 \\ 0.03974367 \\ 0.113747 \\ 0.00862653 \\ 0.00720932 \\ 0.00455974 \\ 0.04984904 \\ 0.00665475 \\ 0.00030809 \\ 0.05292994 \\ 0.02581798 \\ 0.0541623 \\ 0.07246288 \\ 0.02427753 \\ 0.00770226 \\ 0.0592766 \\ 0.0657465 \\ 0.03561526 \\ 0.0337051 \\ 0.00591534 \\ 0.00012324 \\ 0.00252634 \\ 0.00788712 \\ 0.0027112 \\ 0.16815577 \end{bmatrix}$$

4. Treat e10.txt as a test document x. Represent x as a bag-of-words count vector (Hint: the vocabulary has size 27). Print the bag-of-words vector x.

$$x = \begin{bmatrix} 164 \\ 32 \\ 53 \\ 57 \\ 311 \\ 55 \\ 51 \\ 140 \\ 140 \\ 3 \\ 6 \\ 85 \\ 64 \\ 139 \\ 182 \\ 53 \\ 3 \\ 141 \\ 186 \\ 225 \\ 65 \\ 31 \\ 47 \\ 4 \\ 38 \\ 2 \\ 498 \end{bmatrix}$$

5. Compute $\hat{p}(x \mid y)$ for $y = e, j, s$ under the multinomial model assumption, respectively. Use the formula

$$\hat{p}(x \mid y) = \prod_{i=1}^{d} \theta_{i,y}^{x_i}$$

where $x = (x_1, \ldots, x_d)$. Show the three values: $\hat{p}(x \mid y = e), \hat{p}(x \mid y = j), \hat{p}(x \mid y = s)$. Hint: you may notice that we omitted the multinomial coefficient. This is ok for classification because it is a constant w.r.t. y.

$\hat{p}(x \mid y = e) = 10^{-3405.644556038407}$
$\hat{p}(x \mid y = j) = 10^{-3804.210716450759}$
$\hat{p}(x \mid y = s) = 10^{-3670.8233803708918}$

6. Use Bayes rule and your estimated prior and likelihood, compute the posterior $\hat{p}(y \mid x)$. Show the three values: $\hat{p}(y = e \mid x), \hat{p}(y = j \mid x), \hat{p}(y = s \mid x)$. Show the predicted class label of x.

$\hat{p}(y = e \mid x) = \frac{\hat{p}(x|y=e)p(y=e)}{\sum_{z=e,j,s} \hat{p}(x|y=z)p(z)}$

$= \frac{\hat{p}(x|y=e)p(y=e)}{\hat{p}(x|y=e)p(y=e)+\hat{p}(x|y=j)p(y=j)+\hat{p}(x|y=s)p(y=s)}$

using $p(y = e) = p(y = j) = p(y = s)$, cancel out this from nominator and denominator.
$= \frac{\hat{p}(x|y=e)}{\hat{p}(x|y=e)+\hat{p}(x|y=j)+\hat{p}(x|y=s)}$

$= \frac{\hat{p}(x|y=e)}{\hat{p}(x|y=e)(\frac{\hat{p}(x|y=e)}{\hat{p}(x|y=e)}+\frac{\hat{p}(x|y=j)}{\hat{p}(x|y=e)}+\frac{\hat{p}(x|y=s)}{\hat{p}(x|y=e)})}$

$= \frac{1}{(1+\frac{10^{-3804.210716450759}}{10^{-3405.644556038407}}+\frac{10^{-3670.8233803708918}}{10^{-3405.644556038407}})}$
≈ 1 (Ignoring the last two terms in comparison to 1.)

$\hat{p}(y = j \mid x) = \frac{\hat{p}(x|y=j)p(y=j)}{\hat{p}(x|y=e)p(y=e)+\hat{p}(x|y=j)p(y=j)+\hat{p}(x|y=s)p(y=s)}$

$= \frac{\hat{p}(x|y=j)}{\hat{p}(x|y=j)(\frac{\hat{p}(x|y=e)}{\hat{p}(x|y=j)}+\frac{\hat{p}(x|y=j)}{\hat{p}(x|y=j)}+\frac{\hat{p}(x|y=s)}{\hat{p}(x|y=j)})}$

$= \frac{1}{(\frac{10^{-3405.644556038407}}{10^{-3804.210716450759}}+\frac{10^{-3804.210716450759}}{10^{-3804.210716450759}}+\frac{10^{-3670.8233803708918}}{10^{-3804.210716450759}})}$

$= \frac{1}{10^{(3804.210716450759-3405.644556038407)}+1+10^{(3804.210716450759-3670.8233803708918)}} \approx 0$

$\hat{p}(y = s \mid x) = \frac{\hat{p}(x|y=s)}{\hat{p}(x|y=s)(\frac{\hat{p}(x|y=e)}{\hat{p}(x|y=s)}+\frac{\hat{p}(x|y=j)}{\hat{p}(x|y=s)}+\frac{\hat{p}(x|y=s)}{\hat{p}(x|y=s)})}$

$= \frac{1}{10^{(3670.8233803708918-3405.644556038407)}+10^{(3670.8233803708918-3804.210716450759)}+1} \approx 0$

predicted class label of x is 'e' because $\hat{p}(y = e \mid x)$ is maximum of all.

7. Evaluate the performance of your classifier on the test set (files 10.txt to 19.txt in three languages). Present the performance using a confusion matrix. A confusion matrix summarizes the types of errors your classifier makes, as shown in the table below. The columns are the true language a document is in, and the rows are the classified outcome of that document. The cells are the number of test documents in that situation. For example, the cell with row = English and column = Spanish contains the number of test documents that are really Spanish, but misclassified as English by your classifier.

	English	Spanish	Japanese
English	10	0	0
Spanish	0	10	0
Japanese	0	0	10

8. Take a test document. Arbitrarily shuffle the order of its characters so that the words (and spaces) are scrambled beyond human recognition. How does this shuffling affect your Naive Bayes classifier's prediction on this document? Explain the key mathematical step in the Naive Bayes model that justifies your answer.

   The shuffling of characters won't have any effect on the classification result and the predicted label remains the same. In Naive Bayes classifier, we assume conditional independence of features x_i, and hence we can just consider the count of each character for Naive Bayes classifier's prediction on the given document. Position doesn't affect the count feature of each character. Additionally, We use Bag-of-Words (BoW) representation, in which conditional independence of features holds true and shuffling the document doesn't change the BoW representation.

4 Weka (10 pts)

Perform multinomial Naive Bayes classification. Suggested key steps:

- We want you to experience Weka's TextDirectoryLoader. For this, you need to prepare our documents such that each character becomes a word. The easiest way is to insert a space between characters, but turn original space into the word "space". For example, the document "is the sun dying" should be turned into "i s space t h e space s u n space d y i n g". Then, create 3 subdirectories e, j, s and place each of the 20 documents in that language into the corresponding subdirectory.
- Find out how to ask Weka to use TextDirectoryLoader to load those 60 documents and use the subdirectory name as the class name. This happens in the Preprocess tab in Weka Explorer.
- Apply Filter: filters/unsupervised/Attribute/StringtoWordVec. Click to see options, set outputWordCounts to True (otherwise Weka uses binary absence/presence features, which is less interesting for our purpose). You should see @@class@@ and 27 features.
- Choose Edit... A big document by feature table will show up, where you should see word counts roughly in the range 1–100. Right click on @@class@@, select "Attribute as class". You may save it as an arff file. The tri-color histogram for different features is interesting.
- From the Classify tab, choose Classifier / bayes / NaiveBayesMultinomial
- Let Weka perform 10-fold cross validation. Report the resulting confusion matrix.

confusion matrix:

	English	Spanish	Japanese
English	20	0	0
Spanish	0	20	0
Japanese	0	0	20