

COMP 6721 Applied Artificial Intelligence (Summer 2023)

Assignment #4: Naïve Bayes Classification

Solutions

Question 1 Assume that a fancy food-store sells wild hand-picked mushrooms from a local farmer. In the store, the mushrooms are labelled as *gourmet*, *good*, or *at-your-own-risk*. The store always keeps the following inventory: 25% of its mushrooms are labeled *gourmet*, 50% are labeled *good*, and 25% are labeled *at-your-own-risk*. Mushrooms labeled as *gourmet* have a 5% chance of being poisonous, a *good* mushroom has a 15% chance of poisoning someone, and a *at-your-own-risk* mushroom has a 25% chance.

If Jim bought a mushroom from the store and was poisoned,

- (a) What is the probability that the mushroom had been labeled *gourmet*?

Given:

$$P(\text{gourmet}) = 0.25$$

$$P(\text{good}) = 0.5$$

$$P(\text{ayor}) = 0.25$$

$$P(\text{poisonous}|\text{gourmet}) = 0.05$$

$$P(\text{poisonous}|\text{good}) = 0.15$$

$$P(\text{poisonous}|\text{ayor}) = 0.25$$

$$C_i \in \{\text{gourmet}, \text{good}, \text{at-your-own-risk}\}$$

$$\begin{aligned} P(\text{poisonous}) &= \sum P(\text{poisonous}|C_i) * P(C_i) \\ &= 0.05 * 0.25 + 0.15 * 0.5 + 0.25 * 0.25 \\ &= 0.15 \end{aligned}$$

$$\begin{aligned} P(\text{gourmet}|\text{poisonous}) &= \frac{P(\text{poisonous}|\text{gourmet}) * P(\text{gourmet})}{P(\text{poisonous})} \\ &= \frac{0.05 * 0.25}{0.15} \\ &= 0.083 \end{aligned}$$

- (b) What is the probability that the mushroom had been labeled *good*?

$$\begin{aligned}P(\textit{good}|\textit{poisonous}) &= \frac{P(\textit{poisonous}|\textit{good}) * P(\textit{good})}{P(\textit{poisonous})} \\&= \frac{0.15 * 0.5}{0.15} \\&= 0.5\end{aligned}$$

- (c) What is the probability that the mushroom had been labeled *at-your-own-risk*?

$$\begin{aligned}P(\textit{ayor}|\textit{poisonous}) &= \frac{P(\textit{poisonous}|\textit{ayor}) * P(\textit{ayor})}{P(\textit{poisonous})} \\&= \frac{0.25 * 0.25}{0.15} \\&= 0.417\end{aligned}$$

Question 2 Assume that Cecilia receives many e-mails from her home town in Klinga, where people speak Klinish. If you do not know Klinish, don't worry. It is a simple language made up of only 1,000 words that all start with the letter "k". A Klinish document may also contain words that do not start with "k", but these are considered out-of-vocabulary words (like a proper name, for example). Jack is trying to help Cecilia sort her Inbox into 3 mail folders (Personal, Work and Promotion). However, Jack does not speak Klinish, so all he has to work from are old e-mails that Cecilia has already sorted into the right folders. The table below shows a sample of the data that Jack has gathered from Cecilia's previous e-mails. The table indicates the frequency of each Klinish word in each folder (to be complete, the table should contain 1,000 rows, corresponding to each word in Klinish). For example, the word kiki appeared 30 times in e-mails labelled Personal, 50 times in e-mails about Work, and 9 times in Promotion e-mails.

		Folder		
		Personal	Work	Promotion
Word	kami	45	12	17
	kawa	78	1	67
	keke	0	5	80
	kiki	30	50	9
	koko	6	10	10
	kotuku	5	27	20
	koula	17	56	3
	...			
Total Nb of Words		20,000	25,000	17,000

The table above corresponds to data collected from 50 e-mails labeled *Personal*, 65 e-mails labeled *Work* and 45 e-mails labeled *Promotion*.

Based on the data above, Jack is trying to classify the following two e-mails (note that upper and lower cases should not be distinguished).

Email 1:	Koko kami kawa koula keke
Email 2:	Keke kawa, koko Google koula keke!

- (a) Use a Naïve Bayes classifier without any smoothing, to classify the two e-mails above. Use the sum of logs (base 10), and show the score of each of the 3 classes (Personal, Work and Promotion) and the most likely class.

Solution:

priors:

$$P(\text{Personal}) = 50 / (50 + 65 + 45)$$

$$P(\text{Work}) = 65 / (50 + 65 + 45)$$

$$P(\text{Promotion}) = 45 / (50 + 65 + 45)$$

Email 1: Koko kami kawa koula keke

$$\begin{aligned} \text{score}(\text{Personal}) &= \\ &\log(P(\text{personal})) + \log(P(\text{koko}|\text{personal})) + \log(P(\text{kami}|\text{personal})) + \\ &\log(P(\text{kawa}|\text{personal})) + \log(P(\text{koula}|\text{personal})) + \log(P(\text{keke}|\text{personal})) \\ &= \log(50/160) + \log(6/20,000) + \log(45/20,000) + \log(78/20,000) + \log(17/20,000) + \\ &\log(0/20,000) \\ &= -\infty \end{aligned}$$

$$\begin{aligned} \text{score}(\text{work}) &= \log(P(\text{work})) + \log(P(\text{koko}|\text{work})) + \log(P(\text{kami}|\text{work})) + \\ &\log(P(\text{kawa}|\text{work})) + \log(P(\text{koula}|\text{work})) + \log(P(\text{keke}|\text{work})) \\ &= \log(65/160) + \log(10/25,000) + \log(12/25,000) + \log(1/25,000) + \log(56/25,000) + \\ &\log(5/25,000) \\ &= -17.8546 \end{aligned}$$

$$\begin{aligned} \text{score}(\text{promotion}) &= \log(P(\text{promotion})) + \log(P(\text{koko}|\text{promotion})) + \\ &\log(P(\text{kami}|\text{promotion})) + \log(P(\text{kawa}|\text{promotion})) + \log(P(\text{koula}|\text{promotion})) + \\ &\log(P(\text{keke}|\text{promotion})) \\ &= \log(45/160) + \log(10/17,000) + \log(17/17,000) + \log(67/17,000) + \log(3/17,000) + \\ &\log(80/17,000) \\ &= -15.2664 \end{aligned}$$

highest score is -15.2664 \implies the most likely class is promotion

Email 2: Keke kawa, koko Google koula keke!

notes:

- ignore the word Google.
- keke counts twice

$$\begin{aligned} \text{score}(\text{Personal}) &= \log(P(\text{personal})) + \log(P(\text{keke}|\text{personal})) + \log(P(\text{kawa}|\text{personal})) + \\ &\log(P(\text{koko}|\text{personal})) + \log(P(\text{koula}|\text{personal})) + \log(P(\text{keke}|\text{personal})) \\ &= \log(50/160) + \log(0/2,0000) + \log(78/2,0000) + \log(6/20,000) + \log(17/20,000) + \\ &\log(0/2,0000) \\ &= -\infty \end{aligned}$$

$$\begin{aligned} \text{score}(\text{work}) &= \log(65/160) + \log(5/25,000) + \log(1/25,000) + \log(10/25,000) + \\ &\log(56/25,000) + \log(5/25,000) \end{aligned}$$

$$= -18.2348$$

$$\begin{aligned} \text{score}(\text{promotion}) &= \log(45/160) + \log(80/17,000) + \log(67/17,000) + \log(10/17,000) + \\ &\log(3/17,000) + \log(80/17,000) \\ &= -14.5938 \end{aligned}$$

highest score is -14.5938 \implies the most likely class is promotion

- (b) Do the same as part A above, but this time use “add 0.5 smoothing” (i.e., instead of adding the value 1 to each word frequency, add $\frac{1}{2}$ to each word frequency). Adjust the smoothing formula accordingly, and show all your work. Again, use the sum of logs (base 10), and show the score of each of the 3 classes and the most likely class.

Solution:

		Folder		
		Personal	Work	Promotion
Word	kami	45.5	12.5	17.5
	kawa	78.5	1.5	67.5
	keke	0.5	5.5	80.5
	kiki	30.5	50.5	9.5
	koko	6.5	10.5	10.5
	kotuku	5.5	27.5	20.5
	koula	17.5	56.5	3.5
	...			
Total		20,000	25,000	17,000
Nb of		+0.5 x 1,000	+0.5 x 1,000	+0.5 x 1,000
Words		= 20,500	= 25,500	= 17,500

Email 1: Koko kami kawa koula keke

$$\begin{aligned}
 \text{score}(\text{personal}) &= \log(P(\text{personal})) + \log(P(\text{koko}|\text{personal})) + \log(P(\text{kami}|\text{personal})) + \\
 &\log(P(\text{kawa}|\text{personal})) + \log(P(\text{koula}|\text{personal})) + \log(P(\text{keke}|\text{personal})) \\
 &= \log(50/160) + \log(6.5/20,500) + \log(45.5/20,500) + \log(78.5/20,500) + \\
 &\log(17.5/20,500) + \log(0.5/20,500) \\
 &= -16.7561
 \end{aligned}$$

$$\begin{aligned}
 \text{score}(\text{work}) &= \log(P(\text{work})) + \log(P(\text{koko}|\text{work})) + \log(P(\text{kami}|\text{work})) + \\
 &\log(P(\text{kawa}|\text{work})) + \log(P(\text{koula}|\text{work})) + \log(P(\text{keke}|\text{work})) \\
 &= \log(65/160) + \log(10.5/25,500) + \log(12.5/25,500) + \log(1.5/25,500) + \\
 &\log(56.5/25,500) + \log(5.5/25,500) \\
 &= -17.6373
 \end{aligned}$$

$$\begin{aligned}
 \text{score}(\text{promotion}) &= \log(P(\text{promotion})) + \log(P(\text{koko}|\text{promotion})) + \\
 &\log(P(\text{kami}|\text{promotion})) + \log(P(\text{kawa}|\text{promotion})) + \log(P(\text{koula}|\text{promotion})) + \\
 &\log(P(\text{keke}|\text{promotion})) \\
 &= \log(45/160) + \log(10.5/17,500) + \log(17.5/17,500) + \log(67.5/17,500) + \\
 &\log(3.5/17,500) + \log(80.5/17,500) \\
 &= -15.2227
 \end{aligned}$$

highest score is -15.2227 \implies the most likely class is promotion

		Folder		
		Personal	Work	Promotion
Word	kami	45.5	12.5	17.5
	kawa	78.5	1.5	67.5
	keke	0.5	5.5	80.5
	kiki	30.5	50.5	9.5
	koko	6.5	10.5	10.5
	kotuku	5.5	27.5	20.5
	koula	17.5	56.5	3.5
	...			
Total Nb of Words		20,500	25,500	17,500

Email 2: Keke kawa, koko Google koula keke!

notes:

- ignore the word Google.
- keke counts twice

$$\begin{aligned}
 \text{score}(\text{Personal}) &= \log(P(\text{personal})) + \log(P(\text{keke}|\text{personal})) + \log(P(\text{kawa}|\text{personal})) + \\
 &\log(P(\text{koko}|\text{personal})) + \log(P(\text{koula}|\text{personal})) + \log(P(\text{keke}|\text{personal})) \\
 &= \log(50/160) + \log(0.5/20,500) + \log(78.5/20,500) + \log(6.5/20,500) + \\
 &\log(17.5/20,500) + \log(0.5/20,500) \\
 &= -18.7152
 \end{aligned}$$

$$\begin{aligned}
 \text{score}(\text{work}) &= \log(65/160) + \log(5.5/25500) + \log(1.5/25500) + \log(10.5/25500) + \\
 &\log(56.5/25500) + \log(5.5/25500) \\
 &= -17.9939
 \end{aligned}$$

$$\begin{aligned}
 \text{score}(\text{promotion}) &= \log(45/160) + \log(80.5/17,500) + \log(67.5/17,500) + \\
 &\log(10.5/17,500) + \log(3.5/17,500) + \log(80.5/17,500) \\
 &= -14.5599
 \end{aligned}$$

highest score is -14.5599 \implies the most likely class is promotion

Question 3 Let's write a *Python* program to train and run a model using the *Multinomial Naïve Bayes Classifier*. You can use the implementation provided with *scikit-learn*:¹. How to submit your Python Program: create a zip folder containing (a) your python codes .py and (b) a README.txt file on how to run your codes. Submit the zip file along the pdf document in Moodle.

```
import numpy as np
from sklearn.naive_bayes import MultinomialNB
from sklearn.feature_extraction.text import CountVectorizer
```

Start by implementing the *Email Spam Classifier* you've worked through in the lecture. Create the training data:

```
corpus = np.array([
    'cheap meds for sale',
    'click here for the best meds',
    'book your trip',
    'cheap book sale, not meds',
    'here is the book for you'
])
```

To transform the text corpus into a feature vector, you can use *scikit-learn's CountVectorizer*:²

```
vectorizer = CountVectorizer()
X = vectorizer.fit_transform(corpus)
```

You also need the target vector with the labels for the training data (here, spam is 0 and ham is 1):

```
y = np.array([0,0,0,1,1])
```

Get a classifier using the prior probabilities for each class (0.6 for spam, 0.4 for ham):

```
classifier = MultinomialNB(class_prior=[0.6, 0.4])
```

Train a model using your classifier:

```
model = classifier.fit(X, y)
```

Now you can try to apply your model to classify a new email as SPAM or HAM. Here is the example email ('the cheap book') as a feature vector:

```
new_mail = [[0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0]]
```

¹See https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html

²See https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

- (a) Create a complete, working Python program. Print out the intermediate variables to see the data you are working with. Predict the class for the `new_mail` using your `model` and print it out.

Here is a possible solution:

```
import numpy as np
from sklearn.naive_bayes import MultinomialNB
from sklearn.feature_extraction.text import CountVectorizer

corpus = np.array([
    'cheap meds for sale',
    'click here for the best meds',
    'book your trip',
    'cheap book sale, not meds',
    'here is the book for you'
])

# Create training data
vectorizer = CountVectorizer()
X = vectorizer.fit_transform(corpus)
print(vectorizer.get_feature_names())
print(X.toarray())

# Create target vector
y = np.array([0,0,0,1,1])
print("Target vector = ", y)

# Create multinomial naive Bayes classifier
# with prior probabilities of each class
classifier = MultinomialNB(class_prior=[0.6, 0.4])

# Train a model
model = classifier.fit(X, y)

# New email: 'the cheap book'
new_mail = [[0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0]]

# Predict new observation's class
predict = model.predict(new_mail)
print('Predicted class = ', predict)
```

- (b) Inspect the `scikit-learn` documentation to understand how *smoothing* is handled for this algorithm.

As you can see from the `scikit-learn` documentation,

```
class sklearn.naive_bayes.MultinomialNB(alpha=1.0,
    fit_prior=True, class_prior=None)
```

the parameter `alpha` controls smoothing and uses 1.0 by default (this is what we also did in the lecture and is called *Laplace smoothing*³). Try experimenting with switching it off by setting the value to 0.

- (c) Change the code to transform the `new_mail` automatically from a string into a feature vector.

You can use the existing `CountVectorizer`, but you have to make sure you use `transform` instead of `fit_transform` (look up the difference in the documentation):

```
T = vectorizer.transform(np.array(['the cheap book']))
```

- (d) Change the code to automatically compute the prior probabilities using the training data. Print out the priors for the model to verify that they are indeed correct.

If you do not explicitly provide the class priors, like we did above, they will be automatically computed from the training data. So unless you already know the priors for a specific problem, you would simply create the classifier with:

```
classifier = MultinomialNB()
```

You can verify them by printing out the `model` priors with:

```
print("Class priors ln = ", model.class_log_prior_)
```

which will show you the *natural logarithm* (ln) values of the computed class priors.

³See https://en.wikipedia.org/wiki/Additive_smoothing