DSE 210 Final Exam

Orysya Stus

1. A 2 w 4 r $P(A)=0.5$ $P(B)=P(C)$

B 8 w 5 r

C 1 w 3 r

1. $P(A|red) = \frac{P(A \cap red)}{P(red)} = \frac{P(red|A)P(A)}{P(red)} = \frac{(4/6)(0.5)}{(4/6)(0.5)+(5/13)(0.25)+(3/4)(0.25)} = \frac{0.33}{0.33+0.096+0.1875}$

$\approx \frac{0.33}{0.6135} \approx 0.538$

2. $P(white) = 1 - P(red) = 0.3865$

2a. let the set be $S = \{n_1, n_2, \ldots, n_n\}$

For each element we have 2 ways to make our subset: either choose x or leave it

$$\forall n_i \in S, \quad n_i \begin{cases} \to 1 & \text{present in subset} \\ \to 0 & \text{absent in subset} \end{cases}$$

When we do this for all x_i's,

then $2 \times 2 \times 2, \ldots, \times 2$ n times

↳ then # of subsets of this set is 2^n

b. Pascals identity $\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}$ or $\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1}$

let $n=5$ then

$2^5 - 1 = \binom{5}{5} + \binom{5}{4} + \binom{5}{3} + \binom{5}{2} + \binom{5}{1} = 2^n - 1$

$2^{5-1} = \binom{4}{4} + \binom{4}{3} + \binom{4}{2} + \binom{4}{1} + \binom{4}{0}$

$2^{5-2} = \binom{3}{3} + \binom{3}{2} + \binom{3}{1} + \binom{3}{0}$

$2^{5-3} = \binom{2}{2} + \binom{2}{1} + \binom{2}{0}$

$2^{5-4} = \binom{1}{1} + \binom{1}{0}$

$2^{5-5} = \binom{0}{0}$

$$\sum_{k=1}^{n} 2^{n-k} = \underbrace{\binom{4}{1} + \binom{4}{0}}_{\binom{5}{1}} + \underbrace{\binom{0}{0}}_{\binom{5}{5}} + \underbrace{\binom{4}{2} + \binom{3}{1} + \binom{3}{0}}_{\binom{5}{2}} + \underbrace{\binom{4}{3} + \binom{3}{2} + \binom{2}{1} + \binom{2}{0}}_{\binom{5}{3}}$$

$$+ \underbrace{\binom{4}{4} + \binom{3}{3} + \binom{2}{2} + \binom{1}{0} + \binom{1}{1}}_{\binom{5}{4}}$$

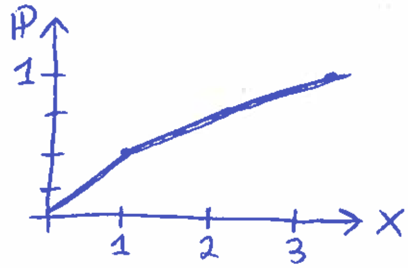
3. $P(X=1)=0.5$ $P(X=2)=0.25$ $P(X=3)=0.25$

1. $E(X) = 1(0.5)+2(0.25)+3(0.25) = 0.5+0.5+0.75 = 1.75$

0. $E(X^2) = 0.5(1^2)+0.25(2^2+3^2) = 0.5+0.25(13) = 3.75$

$var(X) = E(X^2)-(E(X))^2 = 3.75-(1.75^2) = 3.75-3.0625 = 0.6875$

2. plot the CDF of X



4. -

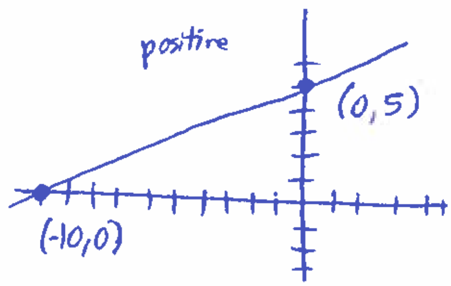
5. $-x+2y \geq 10$

$2y \geq x+10$

$y \geq \frac{1}{2}x+5$

$x=0$ $y \geq 5$

$y=0$ $\frac{1}{2}x \leq -5$ $x \leq -10$



6. A Gaussian PDF $N(0,1)$ $X=2.5$

B " $N(5,2)$ $P(A)=2P(B)$ $P(A)=2/3$ $P(B)=1/3$

~~2. likelihood (A|X=2.5)~~

$\cancel{P(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(X-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(2.5-0)^2}{2(1)^2}\right) = \frac{1}{\sqrt{2\pi}} \exp(\ldots)}$

b. $\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x^2)}{2}\right) = \frac{-2\cdot 1}{\sqrt{8\pi}} \exp\left(-\frac{(x-5)^2}{8}\right)$

$-\frac{x^2}{2} = 0+\left(-\frac{(x-5)^2}{8}\right)$

$-\frac{x^2}{2} = 0-\left(\frac{-\cancel{x}^2-10x+25}{8}\right)$

$\frac{x^2}{2}+\frac{x^2}{8}-\frac{10x}{8} = \frac{-25}{8}$

$-\frac{3}{8}x^2-\frac{5}{4}x = \frac{-25}{8}$

$-\frac{3}{8}x^2-\frac{5}{4}x+\frac{25}{8} = 0$

$x = -5, \frac{5}{3}$

decision boundary if $-5 < X < \frac{5}{3}$ then Um A

$X < -5$ $\frac{5}{3} < X$ then Um B

6a. $P(A|X=2{,}5) = \dfrac{P(2{,}5|A)\,P(A)}{P(x=2{,}5)} = \dfrac{P(X=2{,}5|A)\,P(A)}{P(A)P_A(X=2{,}5)+P(B)P_B(X=2{,}5)}$

$$P(X=2{,}5) = \frac{2}{3}\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-(2{,}5-0)^2}{2}\right) + \frac{1}{3}\frac{1}{\sqrt{2\pi\cdot 4}}\exp\left(-\frac{(2{,}5-5)^2}{2\cdot 4}\right)$$

$$= \frac{1}{3\sqrt{2\pi}}\left(2\exp\left(\frac{-6{,}25}{2}\right) + \frac{1}{2}\exp\left(\frac{-6{,}25}{8}\right)\right) \approx 0{,}04173$$

$P(A) = \frac{2}{3}$

$$P(x=2{,}5|A) = N_A(\mu,\sigma) = \frac{1}{\sqrt{2\pi}}\exp\left(\frac{-6{,}25}{2}\right) \approx 0{,}017528$$

$$P(A|X=2{,}5) \approx \frac{0{,}017528\,(2/3)}{0{,}04173} \approx 0{,}28$$

7. –

8. –

9. $n=1000$ $\mu=307$ stddev $=30$

$$\sigma = \frac{\text{stddev}}{\sqrt{n}} \approx 0.9486$$

∴ national avg $= 307$ w/ stddev ≈ 0.948

0,

	observed men	Women	Total	expected men	women
NM	21 (43.8 %)	9 (16.4%)	30 (29.1%)	14	16
M	20 (41.7 %)	39 (70.9 %)	59 (57.3%)	27	32
W	7 (14.6 %)	7 (12.7%)	14 (13.6 %)	7	7
Total	48	55	103	48	55

Null hypothesis: The 2 sets of #s (for men & women) are independent draws from the same distribution

$$\psi^2 = \sum \frac{[(\text{observed freq.}) - (\text{expected freq})]^2}{\text{expected freq.}}$$

$$\psi^2 = \frac{7^2}{14} + \frac{7^2}{27} + \frac{0^2}{7} + \frac{7^2}{16} + \frac{7^2}{32} + \frac{0^2}{7} \approx 3.5 + 1.8 + 0 + 3.1 + 1.53 \approx 9.93$$

ψ^2 distribution w/ 2 degrees of freedom. A value ≥ 10 has probability roughly 0.01 %

∴ Don't reject null hypothesis

DSE210_Final_Exam

March 20, 2017

1 DSE 201 – Final Exam

1.1 Orysya Stus

1.2 3.20.2017

```
In [50]: import pandas as pd
         import numpy as np
         import sklearn
         from sklearn.naive_bayes import BernoulliNB
         from sklearn.cluster import KMeans
         from sklearn.mixture import GMM
         from sklearn import metrics
         from sklearn import datasets
         import seaborn as sns
         import matplotlib.pyplot as plt
         import re

         %matplotlib inline
```

1.3 For a refresher on Bernoulli Native Bayes review:

http://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.BernoulliNB.html#sklearn.naive_bayes

http://scikit-learn.org/stable/auto_examples/text/document_classification_20newsgroups.html#sphx-glr-auto-examples-text-document-classification-20newsgroups-py

1.4 Q4

For this problem, you'll be using the 20 Newsgroups data set. There are several versions of it on the web. You should download "20news-bydate.tar.gz" from http://qwone.com/~jason/20Newsgroups/

The same website has a processed version of the data, "20news-bydate-matlab.tgz", that is particularly convenient to use. Download this and also the file "vocabulary.txt". Look at the first training document in the processed set and the corresponding original text document to understand the relation between the two. The words in the documents constitute an overall vocabulary V of size 61188. Build a Bernoulli Naive Bayes model using the training data. Write a routine that uses this naive Bayes model to classify a new document. To avoid underflow, work with logs rather than multiplying together probabilities.

1.4.1 Part (a)

Evaluate the performance of your model on the test data. What error rate do you achieve?

In [2]:
```
train = sklearn.datasets.load_files('./20news-bydate-train/', load_content=
test = sklearn.datasets.load_files('./20news-bydate-test/', load_content=Tr
```

In [3]:
```
vocab = {}
reverse_vocab = {}
count = 0
a = open('./vocabulary.txt', 'r')
for v in a:
    val = v.strip()
    vocab[val] = count
    reverse_vocab[count] = val
    count += 1
```

In [4]:
```
categories = {}
for i,e in enumerate(test.target_names):
    categories[i] = e
```

In [5]:
```
test_map = dict(enumerate(test.filenames))
reverse_test_map = {v:k for k,v in test_map.items()}
```

In [6]:
```
from sklearn.feature_extraction.text import CountVectorizer
count_vect = CountVectorizer(strip_accents='unicode', decode_error = 'ignor
X_train_counts = count_vect.fit_transform(train.data)
training_feature_names = count_vect.get_feature_names()
X_train_counts.shape
```

Out[6]: (11314, 61188)

In [7]:
```
from sklearn.feature_extraction.text import TfidfTransformer
tf_transformer = TfidfTransformer()
X_train_tfidf = tf_transformer.fit_transform(X_train_counts)
X_train_tfidf.shape
```

Out[7]: (11314, 61188)

In [8]:
```
count_vect = CountVectorizer(strip_accents='unicode', decode_error = 'ignor
X_test_counts = count_vect.fit_transform(test.data)
test_feature_names = count_vect.get_feature_names()
X_test_counts.shape

tf_transformer = TfidfTransformer()
X_test_tfidf = tf_transformer.fit_transform(X_test_counts)
X_test_tfidf.shape
```

Out[8]: (7532, 61188)

```
In [9]: clf = BernoulliNB(alpha=1.0).fit(X_train_tfidf, train.target)
        predicted = clf.predict(X_test_tfidf)
```

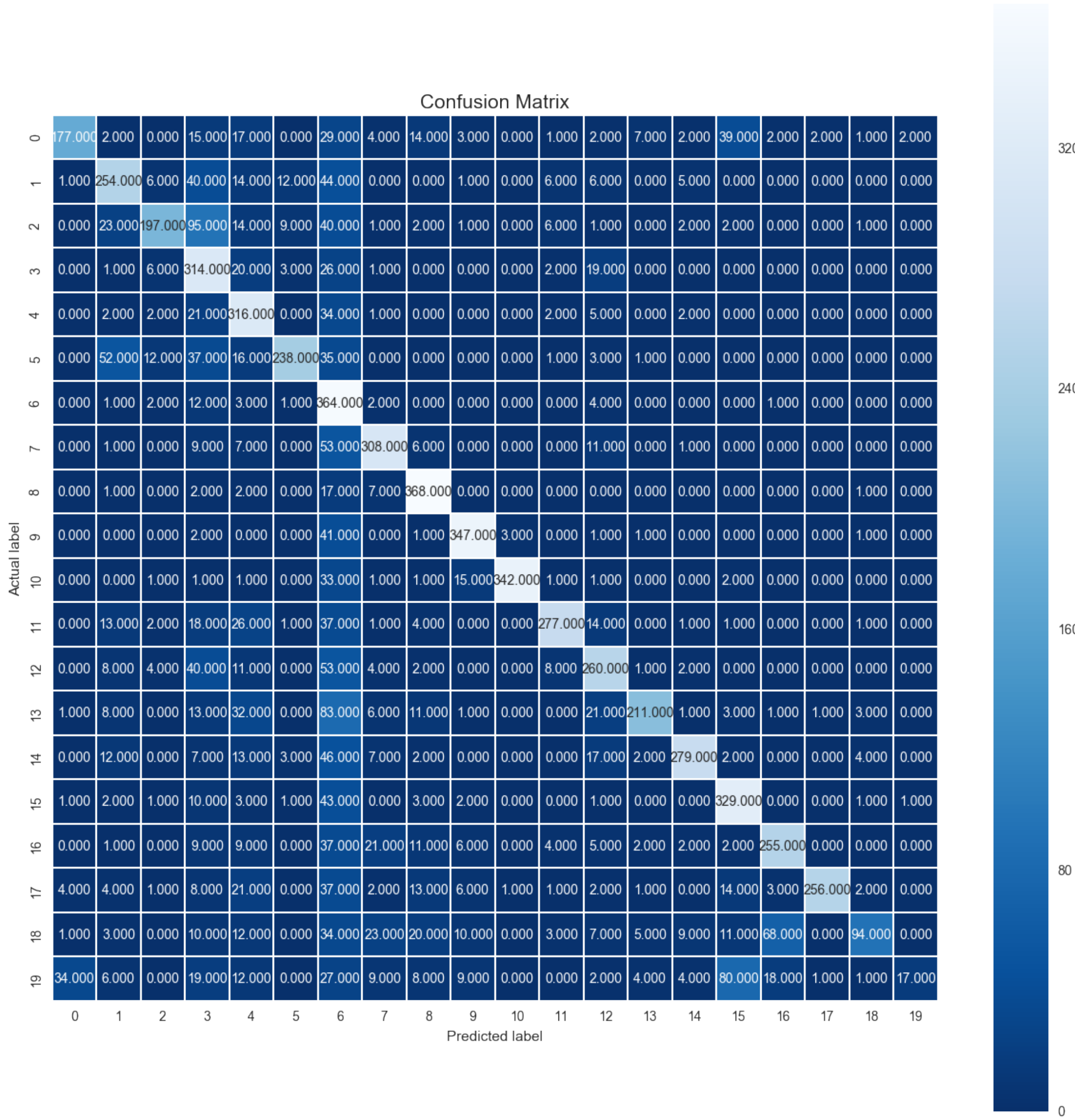
```
In [10]: print 'The model is', metrics.accuracy_score(test.target, predicted), 'acc
         print 'The model has an error rate of', 1- metrics.accuracy_score(test.tar
```

```
The model is 0.690785979819 accurate
The model has an error rate of 0.309214020181
```

1.4.2 Part (b)

Evaluate your final model on the test data. Construct a confusion matrix.

```
In [11]: cm = pd.DataFrame(metrics.confusion_matrix(test.target, predicted))
         plt.figure(figsize=(15, 15))
         sns.heatmap(cm, annot=True, fmt=".3f", linewidth=.5, square = True, cmap =
         plt.ylabel('Actual label');
         plt.xlabel('Predicted label');
         plt.title('Confusion Matrix', size = 15);
```

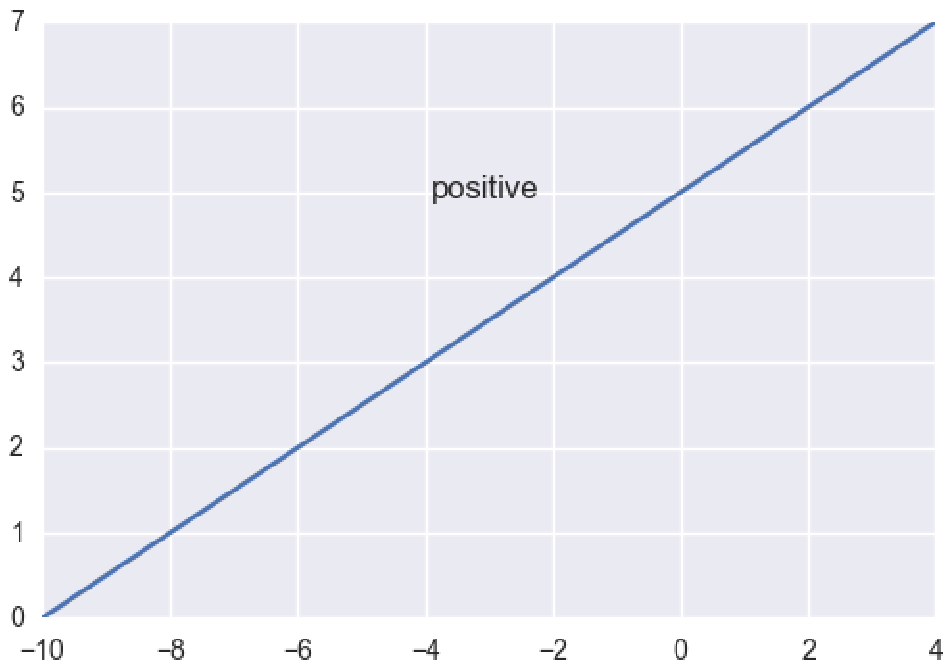


1.5 Q5

```
In [21]: x = np.linspace(-10, 4, 10, endpoint=True)
         y = (1 * x + 10) / 2
         fig = plt.figure()
         ax = fig.add_subplot(111)
         ax.plot(x, y)
         ax.spines['right'].set_color('none')
         ax.spines['top'].set_color('none')
         ax.xaxis.set_ticks_position('bottom')
         ax.spines['bottom'].set_position(('data',0))
```

```
ax.yaxis.set_ticks_position('left')
plt.annotate(r'positive', xy=(-6, 4), xycoords='data', xytext=(+50, +30),
```

Out[21]: <matplotlib.text.Annotation at 0x15df7278>



1.6 Q7

Consider the following observations:

X = (-0.1, -0.2, 0.1, 0.2, 0, 0.1, -0.1, 0, -0.05, 0.1, 1.05, 1.1, 0.9, 0.8, 0.9, 1, 1.2, 1.1, 1.2, 0.9)
Cluster this data into 2 classes using the K-means algorithm. What are the cluster center?
http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

In [13]:
```
X = [-0.1, -0.2, 0.1, 0.2, 0, 0.1, -0.1, 0, -0.05, 0.1, 1.05, 1.1, 0.9, 0.
X = np.array(X)
kmeans = KMeans(n_clusters=2, random_state=0)
kmeans.fit(X.reshape(-1,1))
```

Out[13]:
```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
    n_clusters=2, n_init=10, n_jobs=1, precompute_distances='auto',
    random_state=0, tol=0.0001, verbose=0)
```

In [14]:
```
kmeans.cluster_centers_
```

Out[14]:
```
array([[ 1.015],
       [ 0.005]])
```

1.7 Q8

Construct a Gaussian Mixture Model for the above data. What is your estimate for the number of mixtures?

http://www.astroml.org/book_figures/chapter4/fig_GMM_1D.html
http://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html#sklearn.mixture.

In [56]:
```
import warnings
warnings.filterwarnings('ignore')

for i in xrange(10):
    gmm = GMM(n_components=i+1)
    gmm.fit(X.reshape(-1,1))
    print 'Gaussian mixture model for', i+1, 'components, akaike informati
    print 'Gaussian mixture model for', i+1, 'components, bayesian informa
    print 'Gaussian mixture model for', i+1, 'components, means or estimat
```

```
Gaussian mixture model for 1 components, akaike information criterion (aic) for the
Gaussian mixture model for 1 components, bayesian information criterion (bic) for t
Gaussian mixture model for 1 components, means or estimates are:
[[ 0.51]]
Gaussian mixture model for 2 components, akaike information criterion (aic) for the
Gaussian mixture model for 2 components, bayesian information criterion (bic) for t
Gaussian mixture model for 2 components, means or estimates are:
[[ 1.015]
 [ 0.005]]
Gaussian mixture model for 3 components, akaike information criterion (aic) for the
Gaussian mixture model for 3 components, bayesian information criterion (bic) for t
Gaussian mixture model for 3 components, means or estimates are:
[[ 0.99375351]
 [ 0.005     ]
 [ 1.02049175]]
Gaussian mixture model for 4 components, akaike information criterion (aic) for the
Gaussian mixture model for 4 components, bayesian information criterion (bic) for t
Gaussian mixture model for 4 components, means or estimates are:
[[ 0.99765427]
 [ 0.00269455]
 [ 1.02144721]
 [ 0.01174379]]
Gaussian mixture model for 5 components, akaike information criterion (aic) for the
Gaussian mixture model for 5 components, bayesian information criterion (bic) for t
Gaussian mixture model for 5 components, means or estimates are:
[[ 0.0100255 ]
 [ 1.01030616]
 [ 0.00269286]
 [ 0.9902593 ]
 [ 1.02263586]]
Gaussian mixture model for 6 components, akaike information criterion (aic) for the
Gaussian mixture model for 6 components, bayesian information criterion (bic) for t
```

```
Gaussian mixture model for 6 components, means or estimates are:
[[ 1.02250557]
 [ 0.00576591]
 [ 0.99368412]
 [ 0.01246569]
 [ 1.01279667]
 [ 0.00228867]]
Gaussian mixture model for 7 components, akaike information criterion (aic) for the
Gaussian mixture model for 7 components, bayesian information criterion (bic) for t
Gaussian mixture model for 7 components, means or estimates are:
[[ 0.00189   ]
 [ 1.01307606]
 [ 0.99493135]
 [ 0.00571948]
 [ 0.0161848 ]
 [ 0.01059467]
 [ 1.02303845]]
Gaussian mixture model for 8 components, akaike information criterion (aic) for the
Gaussian mixture model for 8 components, bayesian information criterion (bic) for t
Gaussian mixture model for 8 components, means or estimates are:
[[ 0.99831262]
 [ 0.00780372]
 [ 1.01345556]
 [ 0.00193757]
 [ 0.01177073]
 [ 0.00493955]
 [ 1.02313641]
 [ 0.01784174]]
Gaussian mixture model for 9 components, akaike information criterion (aic) for the
Gaussian mixture model for 9 components, bayesian information criterion (bic) for t
Gaussian mixture model for 9 components, means or estimates are:
[[ 0.01164422]
 [ 1.01509937]
 [ 0.00778792]
 [ 1.00085625]
 [ 0.00459122]
 [ 1.02451423]
 [ 0.00179331]
 [ 0.01629944]
 [ 0.99233726]]
Gaussian mixture model for 10 components, akaike information criterion (aic) for th
Gaussian mixture model for 10 components, bayesian information criterion (bic) for
Gaussian mixture model for 10 components, means or estimates are:
[[ 0.00447655]
 [ 1.01746192]
 [ 0.01119657]
 [ 0.99794265]
 [ 0.00720346]
```

```
 [ 1.02433142]
 [ 0.00130828]
 [ 1.01137019]
 [ 0.98993299]
 [ 0.01605409]]
```

1.7.1 2 components were used like K-means, showing that the Gaussian mixture model means or estimates are the same as those found through K-means clustering. Also, n_components=2 has the lowest bic and aic therefore providing the best means.

```
In [ ]:
```