Data Mining
Home work 09
Machine Learning Part:2

Aqeel Labash
Lecturer: Jaak Vilo

06 April 2016

First Question

The article focus on classification. And here is the list

- No matter what algorithm you pick it's consist from three parts : Representation, Evaluation, Optimization.
- **Representation:** The feature we want to use in away computer can handle.
- **Evaluation or objective or scoring function:** used to distinguish between good and bad classifiers.
- **Optimization :** a method that search for best scoring classifiers.
- Generalization is what we want. "if there are 100,000 words in the dictionary, the spam filter described above has 2100,000 possible different inputs."
- Always test your module on data different than the train data.
- data by itself is not enough for generalization knowledge is also required to know which module to apply."A corollary of this is that one of the key criteria for choosing a representation is which kinds of knowledge are easily expressed in it. For example, if we have a lot of knowledge about what makes examples similar in our domain, instance-based methods may be a good choice. If we have knowledge about probabilistic dependencies, graphical models are a good fit.And if we have knowledge about what kinds of preconditions are required by each class, "IF . . . THEN . . ." rules may be the best option."
- **Over fitting** could be decomposed to Bias and Variance."Bias is a learner's tendency to consistently learn the same wrong thing.Variance is the tendency to learn random things irrespective of the real signal."
- Some ways to fight over fitting is: regularization term, cross validation, statistical significance.
- In most cases over fitting doesn't happen because of noise.
- intuitions used for low dimension usually don't work with high dimension problems."our intuitions, which come from a three-dimensional world, often do not apply in high-dimensional ones."
- Numbers in theory might not be applicable and not always correct."consider the space of Boolean functions of d Boolean variables. If there are e possible different examples, there are 2^e possible different functions, so since there are 2^d possible examples,the total number of functions is 2^{2^d}.And even for hypothesis spaces that are "merely" exponential, the bound is still very loose, because the union bound is very pessimistic. For example, if there are 100 Boolean features and the hypothesis space is decision trees with up to 10 levels, to guarantee $\delta = \epsilon = 1\%$ in the bound above we need half a million examples. But in practice a small fraction of this suffices for accurate learning."
- **Good Features:**are those which independent from each other and highly correlated with the class.
- Awesome feature is not necessarily provided by the data we might have to build it."Often, the raw data is not in a form that is amenable to learning, but you can construct features from it that are."
- The algorithm with more data wins."As a rule of thumb, a dumb algorithm with lots and lots of data beats a clever one with modest amounts of it. (After all, machine learning is all about letting data do the heavy lifting.)"

- The more data we have, the more complex the classifier.
- "Variable size learners can in principle learn any function given sufficient data, but in practice they may not, because of limitations of the algorithm (for example,greedy search falls into local optima) or computational cost."
- Ensemble learning give better results usually.
- if the classifier is simpler doesn't mean it's more accurate.
- being able to represent the data doesn't mean that module can learn it."For example, standard decision tree learners cannot learn trees with more leaves than there are training examples."
- correlation doesn't mean one cause another.it's just an observation on the data that might have cause relation.

Second Question

For this task I used the following Code :

```
1
2  import matplotlib.pyplot as plt
3
4
5  # In[2]:
6
7  original = {}
8  TotalPositive = 0
9  TotalNegative = 0
10 with open('data.class','r') as f:
11 f = f.readlines()
12 for line in f:
13 line = line.split()
14 if line[1]=='T':
15 original[line[0]]=True
16 TotalPositive+=1
17 else:
18 original[line[0]]=False
19 TotalNegative+=1
20 roc1=[]
21 with open('roc1.txt','r') as f:
22 f = f.readlines()
23 for line in f:
24 roc1.append(line.strip())
25 roc2=[]
26 with open('roc2.txt','r') as f:
27 f = f.readlines()
28 for line in f:
29 roc2.append(line.strip())
30 roc3=[]
31 with open('roc3.txt','r') as f:
32 f = f.readlines()
33 for line in f:
34 roc3.append(line.strip())
35 roc4=[]
36 with open('roc4.txt','r') as f:
37 f = f.readlines()
38 for line in f:
39 roc4.append(line.strip())
40 rocperfect = []
41 for x in original.keys():
42 if original[x]:
43 rocperfect = [x]+rocperfect
44 else:
45 rocperfect.append(x)
46
47
48 # In[3]:
49
50 def GetTPFP(k,dataset):
51 TP=0
52 FP=0
53 TN=0
54 FN=0
```

```
for i in range (3000):
#Identified True
if i<k:
TP+=original[dataset[i]]
else:
TN+=original[dataset[i]]
#(TP,FP,TN,FN)
#(F11,F01,F10,F00)
#return (TP,TotalPositive-TP,TN,TotalNegative-TN)
return (float (TP)/float(TotalPositive),float(k-TP)/float(TotalNegative))



# In[4]:

roc1cm=[]
roc2cm=[]
roc3cm=[]
roc4cm=[]
rocperfectcm=[]
print 'processing roc1'
for i in range (3000):
roc1cm.append(GetTPFP(i,roc1))
print 'processing roc2'
for i in range (3000):
roc2cm.append(GetTPFP(i,roc2))
print 'processing roc3'
for i in range (3000):
roc3cm.append(GetTPFP(i,roc3))
print 'processing roc4'
for i in range (3000):
roc4cm.append(GetTPFP(i,roc4))
print 'processing rocperfect'
for i in range (3000):
rocperfectcm.append(GetTPFP(i,rocperfect))
print 'Finished'


# In[5]:

plt.figure('roc1.jpg')
plt.plot([x[1] for x in roc1cm],[x[0] for x in roc1cm], 'ro')
plt.ylabel('TPR')
plt.xlabel('FPR')
plt.title('Roc1')
#plt.show()
plt.savefig('roc1.jpg')


# In[6]:

plt.figure('roc2.jpg')
plt.plot([x[1] for x in roc2cm],[x[0] for x in roc2cm], 'ro')
plt.ylabel('TPR')
plt.xlabel('FPR')
plt.title('Roc2')
#plt.show()
plt.savefig('roc2.jpg')


# In[7]:

plt.figure('roc3.jpg')
plt.plot([x[1] for x in roc3cm],[x[0] for x in roc3cm], 'ro')
plt.ylabel('TPR')
plt.xlabel('FPR')
plt.title('Roc3')
#plt.show()
plt.savefig('roc3.jpg')


# In[8]:

plt.figure('roc4.jpg')
plt.plot([x[1] for x in roc4cm],[x[0] for x in roc4cm], 'ro')
plt.ylabel('TPR')
```

```
plt.xlabel('FPR')
plt.title('Roc4')
#plt.show()
plt.savefig('roc4.jpg')


# In[9]:

plt.figure('rocperfect.jpg')
plt.plot([x[1] for x in rocperfectcm],[x[0] for x in rocperfectcm], 'ro')
plt.ylabel('TPR')
plt.xlabel('FPR')
plt.title('Rocperfect')
#plt.show()
plt.savefig('rocperfect.jpg')


# In[10]:

AUCROC1=0
AUCROC2=0
AUCROC3=0
AUCROC4=0
AUCROCPERFECT=0

for i in range(1,3000):
AUCROC1+= (roc1cm[i][1]-roc1cm[i-1][1])*roc1cm[i][0]
AUCROC2+= (roc2cm[i][1]-roc2cm[i-1][1])*roc2cm[i][0]
AUCROC3+= (roc3cm[i][1]-roc3cm[i-1][1])*roc3cm[i][0]
AUCROC4+= (roc4cm[i][1]-roc4cm[i-1][1])*roc4cm[i][0]
AUCROCPERFECT+= (rocperfectcm[i][1]-rocperfectcm[i-1][1])*rocperfectcm[i][0]

print AUCROC1,AUCROC2,AUCROC3,AUCROC4,AUCROCPERFECT
```

What I did in the previous is simply calculated TP and from it I calculated TPR & FPR.

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN} = 1 - specificity$$

the confusion matrix and from it I calculated TPR,FPR then draw them.And here is the figures :

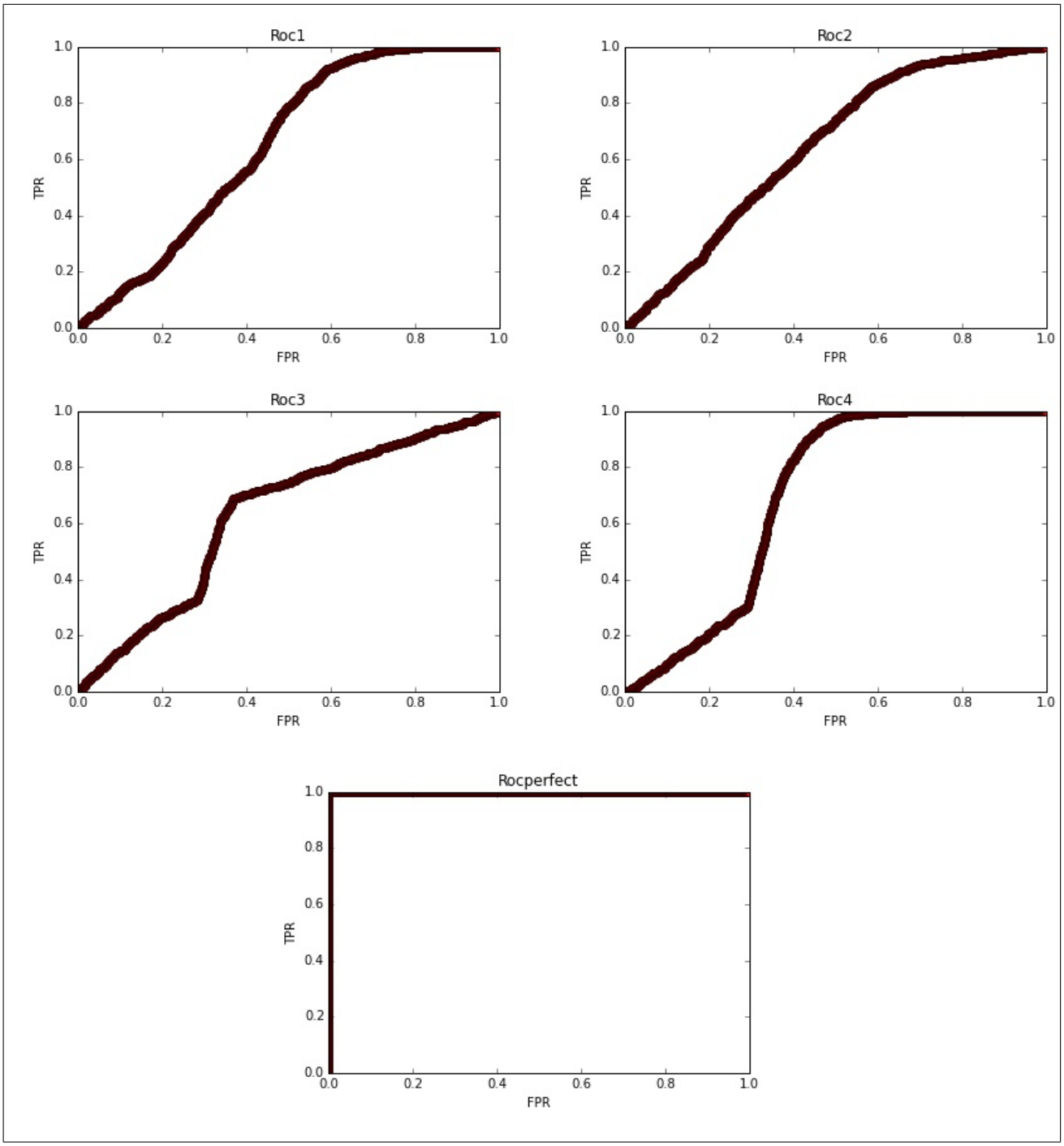

Figure 1: ROC for all the predictions.

We can notice from the previous figures that Roc4 give a good prediction for TP and minimum FP compared to others.

Note:I added a plot for the perfect file how the perfect classifier would look like. The following code calculate approximation of the area under curve.We calculate the sum of all rectangles (point by point) and so on.

```
AUCROC1=0
AUCROC2=0
AUCROC3=0
AUCROC4=0

for i in range(1,3000):
AUCROC1+= (roc1cm[i][1]-roc1cm[i-1][1])*roc1cm[i][0]
AUCROC2+= (roc2cm[i][1]-roc2cm[i-1][1])*roc2cm[i][0]
AUCROC3+= (roc3cm[i][1]-roc3cm[i-1][1])*roc3cm[i][0]
AUCROC4+= (roc4cm[i][1]-roc4cm[i-1][1])*roc4cm[i][0]
print AUCROC1,AUCROC2,AUCROC3,AUCROC4
```

The previous code give the following result :

$$AUC1 = 0.650940277346, AUC2 = 0.647270924831, AUC3 = 0.629960231006, AUC4 = 0.696844993141$$

Third Question

To charactraize the previous plots.I would explain what is happening.When we start we predict 0 or 1 positive so TP+FP =1 at best cases, while k is growing TP+FP is growing as well.At the end we predict all the data as True.In that point we have $TPR = 1, FPR = 1$.
The previous figures For this task I used the hint for this question so I subtracted FPR from TPR and then tried to find the max value. I used the following code :

```
def FindK(dataset):
k=0
lst=[]
for i in range (3000):
lst.append({'x':i,'y':dataset[i][0]-dataset[i][1]})

lst.sort(key=lambda x: x['y'], reverse=True)
return lst[0]['x']
print FindK(roc1cm),FindK(roc2cm),FindK(roc3cm),FindK(roc4cm)
roc1best = FindK(roc1cm)
roc2best = FindK(roc2cm)
roc3best = FindK(roc3cm)
roc4best = FindK(roc4cm)
rocperfectbest= FindK(rocperfectcm)
print roc1best,roc2best,roc3best,roc4best,rocperfectbest
```

the previous code output the following data :

$$roc1best : 2179, roc2best : 2089, roc3best : 1500, roc4best : 1996, rocperfectbest : 1215$$

To get better view I drawn the data and pointed out those points using the following code:

```
plt.figure('roc1_best.jpg')
plt.plot([x[1] for x in roc1cm],[x[0] for x in roc1cm], 'ro')
plt.ylabel('TPR')
plt.xlabel('FPR')
plt.title('Roc1 Best Point')
plt.axvline(x=roc1cm[roc1best][1])
plt.axhline(y=roc1cm[roc1best][0])
#plt.show()
plt.savefig('roc1_best.jpg')


# In[24]:

plt.figure('roc2_best.jpg')
plt.plot([x[1] for x in roc2cm],[x[0] for x in roc2cm], 'ro')
plt.ylabel('TPR')
plt.xlabel('FPR')
plt.title('Roc2 Best Point')
plt.axvline(x=roc2cm[roc2best][1])
plt.axhline(y=roc2cm[roc2best][0])
#plt.show()
plt.savefig('roc2_best.jpg')


# In[25]:

plt.figure('roc3_best.jpg')
plt.plot([x[1] for x in roc3cm],[x[0] for x in roc3cm], 'ro')
plt.ylabel('TPR')
plt.xlabel('FPR')
plt.title('Roc3 Best Point')
plt.axvline(x=roc3cm[roc3best][1])
plt.axhline(y=roc3cm[roc3best][0])
#plt.show()
plt.savefig('roc3_best.jpg')


# In[27]:

plt.figure('roc4_best.jpg')
```

```
plt.plot([x[1] for x in roc4cm],[x[0] for x in roc4cm], 'ro')
plt.ylabel('TPR')
plt.xlabel('FPR')
plt.title('Roc4 Best Point')
plt.axvline(x=roc4cm[roc4best][1])
plt.axhline(y=roc4cm[roc4best][0])
#plt.show()
plt.savefig('roc4_best.jpg')


# In[28]:

plt.figure('rocperfect_best.jpg')
plt.plot([x[1] for x in rocperfectcm],[x[0] for x in rocperfectcm], 'ro')
plt.ylabel('TPR')
plt.xlabel('FPR')
plt.title('Rocperfect Best Point')
plt.axvline(x=rocperfectcm[rocperfectbest][1])
plt.axhline(y=rocperfectcm[rocperfectbest][0])
#plt.show()
plt.savefig('rocperfect_best.jpg')
```

The previous code output the following figures :

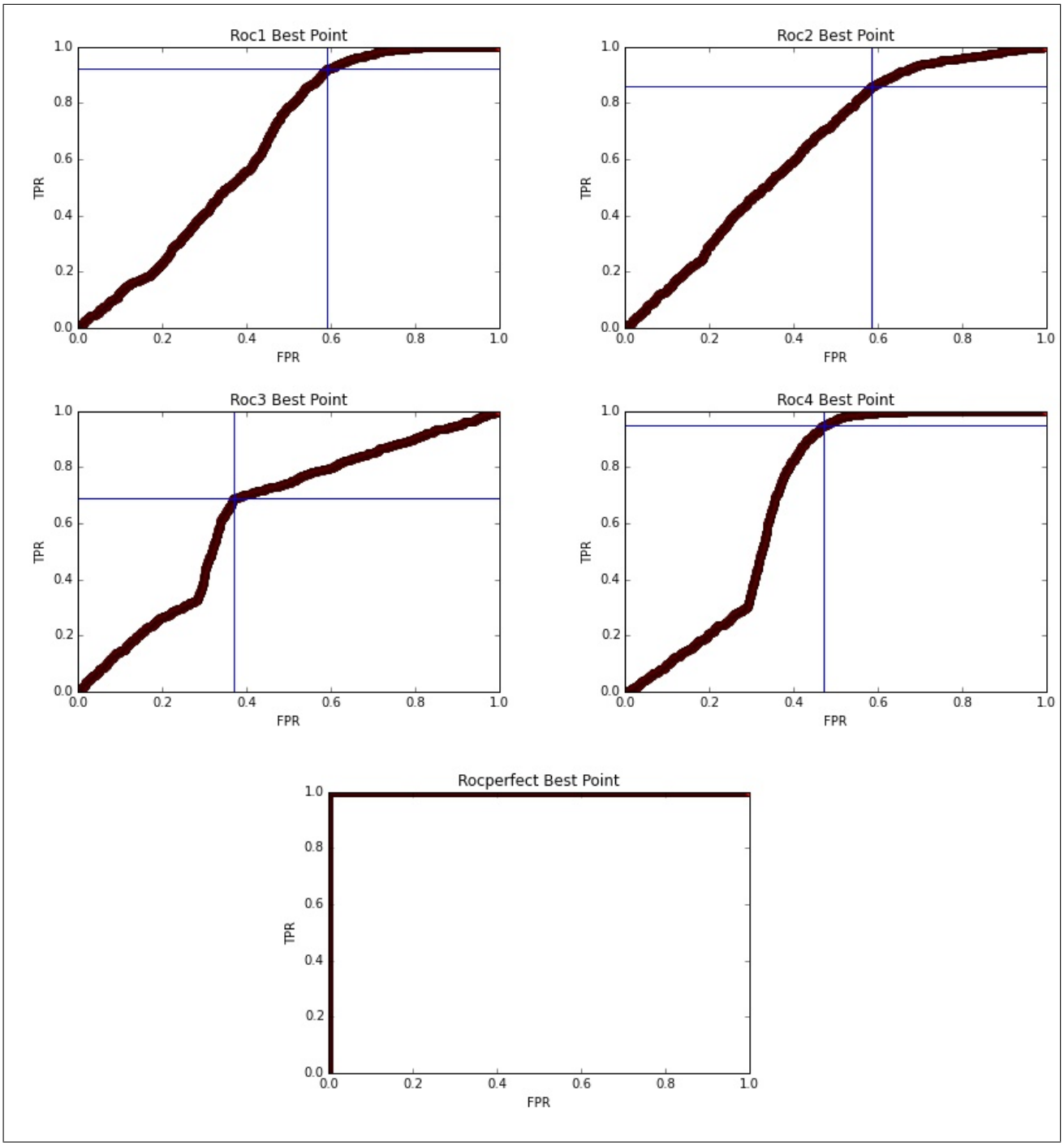

Figure 2: ROC for all the predictions pointing the best split point.

From the previous figure we can see that at the "Rocperfect" we can't see the blue lines that's because they overlap with the data itself.

Fourth Question

Fifth Question

Sixth Question

References

[1] receiver operating Characteristics