Exam INF5830, V2014,
Solutions exercices 1 & 4

1 Evaluation (25%)

(a) Kim is constructing a classifier which reads medical records and determines whether the patient has a certain symptom, call it *SYMP1*, or not. The classifier is tested on a test set of 500 records where 150 of the records are taken to contain *SYMP1*, while the other records do not contain the symptom. Of the 150 records that contain *SYMP1*, the classifier correctly classifies 60 records to contain the symptom, but fails to recognize the other 90 records. Of the 350 records that do not contain the symptom, the classifier says of 40 of them that they do contain the symptom and of 310 of them that they do not contain the symptom. What is the accuracy, recall, precision and F-score of this classifier?

		Classifier	
		SYMP1	not-SYMP1
Reference	SYMP1	60	90
	not-SYMP1	40	310

- The true positives are $tp = 60$
- The false positives are $fp = 40$
- The true negatives are $tn = 310$
- The false negatives are $fn = 90$
- The sample size is $n = 500$
- From its definiton the accuracy equals $\frac{tp+tn}{n} = \frac{60+310}{500} = 0.74$
- The recall equals $\frac{tp}{tp+fn} = \frac{60}{60+90} = 0.4$
- The precision equals $\frac{tp}{tp+fp} = \frac{60}{60+40} = 0.6$
- The F-score equals $\frac{2PR}{P+R} = \frac{2*0.4*0.6}{0.4+0.6} = 0.48$

(b) Kim also wants to check the records for a different symptom, *SYMP2*, and constructs a classifier for this. It is tested on the same 500 records as the first classifier. Of the 50 records that contains *SYMP2*, the classifier recognizes 10 and fails to recognize 40. All the records that do not contain *SYMP2* are correctly classified to not contan the symptom. What is the precision and recall of this classifier?

What are the macroaveraged precision and recall of the two classifiers?

		Classifier	
		SYMP2	not-SYMP2
Reference	SYMP2	10	40
	not-SYMP2	0	450

- The true positives are $tp = 10$
- The false positives are $fp = 0$
- The true negatives are $tn = 450$
- The false negatives are $fn = 40$
- The sample size is $n = 500$
- The recall equals $\frac{tp}{tp+fn} = \frac{10}{10+40} = 0.2$
- The precision equals $\frac{tp}{tp+fp} = \frac{10}{10+0} = 1$
- Macroaveraged recall $= \frac{R1+R2}{2} = \frac{0.4+0.2}{2} = 0.3$
- Macroaveraged precision $= \frac{P1+P2}{2} = \frac{0.6+1.0}{2} = 0.8$

(c) What are the microaveraged recall and precision of the two classifiers.

We make the pooled table:

		Classifier	
		SYMP	not-SYMP
Reference	SYMP	60	130
	not-SYMP	40	800

- The true positives are $tp = 70$
- The false positives are $fp = 40$
- The true negatives are $tn = 760$
- The false negatives are $fn = 130$
- The sample size is $n = 1000$
- The microaveraged recall equals $\frac{tp}{tp+fn} = \frac{70}{70+130} = 0.35$
- The microaveraged precision equals $\frac{tp}{tp+fp} = \frac{70}{70+40} = \frac{7}{11}$

4 Feature selection (25%)

When choosing features for a classification task, one is looking for features which are good discriminators between the classes. We take as example word sense disambiguation and the senses of the word *line* from the hard-line-serve-corpus. Not to make the example too difficult we are only considering two classes, the *product*-class and the *non product*-class. We have selected 500 training items from the *product*-class and 500 training items from the *non product*-class.

There are several different measures which may be used to decide what the best features are, and the different measures might rank candidates differently. This bears similarities to how collocations are ranked. To fix our intuitions, we will take as example two possible features, whether the word *computer* occurs in the context of *line* and whether the word *company* occurs in the context of *line*. The word *computer* occurs in 64 of the 500 instances of the *product*-sense and in 8 of the 500 instances of the *non product*-sense, while the word *company* occurs in 121 of the 500 instances of the *product*-sense and in 41 of the 500 instances of the *non product*-sense.

(a) What does it mean to rank the candidate features by absolute frequencies? Show how absolute frequencies will evaluate the two features. Which of the two features will be ranked highest by this measure?

Alternative 1 There are two alternative solutions, which both will be counted as correct. In the NLTK-chapter on classification we chose the features with the highest total frequency, irrespectively of class. With this counting *computer* occurs $64 + 8 = 72$ giving a score of .072, while *company* occurs $121 + 41 = 162$ giving a score of .162. *Company* is ranked as the best feature.

Alternative 2 In Manning and Schütze, absolute frequencies for collactions were measured by how often the collocation occured. Transfered to this task, it could be interpreted as how often the term occurs in the context of the *product-sense*. With this counting, *computer* will get the score $64/1000 = .064$, and *company* gets the score $121/1000 = .121$. The ranking is the same as with alternative 1.

(b) What does it mean to rank the candidate features by pointwise mutual information? Show how pointwise mutual information will evaluate the two features. Which of the two features will be ranked highest by this measure?

Pointwise mutual information compares the probability (estimated by frequency) of seeing the two phenomena together, compared to probability if the two were independent.

$$\text{PMI}(\text{computer}, \text{prod-sense}) =$$

$$\frac{P(\text{computer}, \text{prod-sense})}{P(\text{computer})P(\text{prod-sense})} = \frac{\frac{64}{1000}}{\frac{72}{1000} * \frac{500}{1000}} = \frac{64}{36} = \frac{16}{9}$$

$$\text{PMI}(\text{company}, \text{prod-sense}) =$$

$$\frac{P(\text{company}, \text{prod-sense})}{P(\text{company})P(\text{prod-sense})} = \frac{\frac{121}{1000}}{\frac{162}{1000} * \frac{500}{1000}} = \frac{121}{81}$$

Since $\frac{16}{9} = \frac{144}{81}$ *computer* is ranked as the best feature.

(c) What does it mean to rank the candidate features by the t-test (or by the Z-test)? You may choose the t-test or Z-test and show how the two features are evaluated by the measure. Which of the two features will be ranked highest by this measure?

In general the Z-test assumes a population with a normal distribution $N(\mu, \sigma)$ and considers a simple random sample X and asks for the probability that X is selected from the population by the formula

$$z = \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

where n is the size of the sample X. We consider the current task a Bernoulli trial where each occurrence of the *product-sense* is an experiment where the outcome is 1 if the feature is present and 0 if it is not present.

The population (corresponding to the null assumption) is that the occurence of the feature and the sense are independent, hence

$$\mu = P(\text{feature}) * P(\text{prod-sense})$$

This is a binomial distribution where

$$p = \mu$$

and where

$$\sigma = \sqrt{p(1-p)}$$

The sample are the actual co-occurrences of the feature and the sense hence

$$\bar{x} = P(\text{feature}, \text{prod-sense})$$

For the two features in question we get

Computer

$$p = \mu = P(\text{computer}) * P(\text{prod-sense}) = .072 * 0.5 = .036$$

$$\sigma = \sqrt{p(1-p)} = \sqrt{.036 * (1 - 0.036)} \approx \sqrt{0.036} = 0.6$$

$$\bar{x} = P(\text{computer}, \text{prod-sense}) = .064$$

$$z = \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{0.064 - 0.036}{\sqrt{\frac{0.036}{1000}}} = \frac{28}{6} = \frac{14}{3}$$

Company

$$p = \mu = P(\text{company}) * P(\text{prod-sense}) = .162 * 0.5 = .081$$

$$\sigma = \sqrt{p(1-p)} = \sqrt{.081 * (1 - 0.081)} \approx \sqrt{0.081} = 0.9$$

$$\bar{x} = P(\text{company}, \text{prod-sense}) = .121$$

$$z = \frac{\bar{x} - \mu}{\sqrt{\frac{\sigma^2}{n}}} = \frac{0.121 - 0.081}{\sqrt{\frac{0.081}{1000}}} = \frac{40}{9}$$

Since $\frac{14}{3} = \frac{42}{9}$, *company* is ranked highest.

END