Reflections on Bayesian Spam Filtering

- Tutorial nr.10 of CS2013 is based on Rosen, 6th Ed., Chapter 6 & exercises
- The following notes discuss why the Bayesian approach was useful in this case, and what might have been done differently

Bayesian Spam filtering

- Task: predict whether a given email is spam
- But what does that mean?
  - » an unseen email?
  - » an email in a specific corpus?
- The standard distinction in statistics between
  - » a sample space (the entire "population" of interest)
  - » a particular sample

Bayesian Spam filtering

- Task: predict whether a given email is spam

- First: focus on one word w (e.g., w = "friend")
- Define P(E) = prob that a random email contains w at least once
- Define P(S) = prob that a random email is spam

- Task: estimate P(S|E)
  - » I read the word "friend" in an email. Is this spam?

Task: estimate P(S|E)

- One task: You have a large corpus C of emails. Your task is to estimate P(S|E) over C
  - » Variant task: estimate P(S|E) for an unseen email
- Assume all emails in your corpus have been marked as Good (G, no spam) or Bad (B, spam)
  - » Variant assumption: you have only had resources to mark a part of the corpus
- Direct approach: compute P(S|E) using frequency: $|C \cap \text{contain } w| \,/\, |C|$
  - » have to compute a different set for each w
  - » C may not be representative (i.e., may not tell you much about an unseen email message)

Task: estimate P(S|E)

- Promising approach: use Bayes' Theorem
  $P(S|E) = P(E|S) * P(S) / P(E)$
- Data:
  $|C| = 1{,}000{,}000$
  $|G| = 20{,}000$, $|B| = 10{,}000$ (only a small part of C!)
  $|G \cap \text{contain } w| = 50$
  $|B \cap \text{contain } w| = 2{,}500$
- Let's first use the approach suggested in the Practical, then explore alternatives

Task: estimate P(S|E)

$P(S|E) = P(E|S) * P(S) / P(E)$

$|C| = 1{,}000{,}000{,}000$

$|B| = 20{,}000$, $|G| = 10{,}000$ (only a small part!)

$|G \cap \text{contain } w| = 50$

$|B \cap \text{contain } w| = 2{,}500$

P(E|S) = 2,500 / 20,000 = 0.125

Assume **P(S) = 0.1** (why not compute P(S)= 20,000/30,000 = 0.66?)

How to estimate **P(E)** ?

Task: estimate P(S|E)

$P(S|E) = P(E|S) * P(S) / P(E)$

$|C| = 1{,}000{,}000{,}000$

$|B| = 20{,}000$, $|G| = 10{,}000$ (only a small part!)

$|G \cap \text{contain } w| = 50$

$|B \cap \text{contain } w| = 2{,}500$

P(E|S) = 2,500 / 20,000 = 0.125

Assume **P(S) = 0.1**

How to estimate **P(E)** ?

(Assume we do not know $|C \cap \text{contain } w|$)

Task: estimate P(S|E)

P(E|S) = 0.125

P(S) = 0.1

How to estimate **P(E)** ?

P(E) =Marginalisation=

P(E,S) + P(E,not-S) =ProductRule=

P(S)*P(E|S) + P(not-S)*P(E|not-S) =

(0.1*0.125) + (0.9*50/10,000) =

0.0125+(0.9*0.005) = 0.0125+0.0045=0.0170

Task: estimate P(S|E)

We can now estimate P(S|E) using Bayes' Theorem:

P(S|E) = P(E|S) * P(S) / P(E)

P(E|S) = 0.125

P(S) = 0.1

P(E) = 0.0170

Prediction: P(S|E) =

0.125*0.1 / 0.0170 =approx=

0.735 (Given the threshold used in the Tutorial, the message is not classified as Spam)

Task: estimate P(S|E)

P(S|E) = P(E|S) * P(S) / P(E)

Some alternatives:

1. (see Practical) You know nothing about P(S).

➔ S is Boolean, hence estimate P(S)=0.5

This is higher than our 0.1, hence you'd be over-estimating the probability P(S|E)

Task: estimate P(S|E)

P(S|E) = P(E|S) * P(S) / P(E)

Some alternatives:

2. You're able to obtain P(E) from $C \cap$ contain w

➔ Great! If *C* is large and representative enough, this may give you a better assessment of P(E) (based on much more data than the estimate above, which estimated P(E|S) and P(E|not-S) on the basis of only 20,000 and 10,000 messages)

Task: estimate P(S|E)

P(S|E) = P(E|S) * P(S) / P(E)

Some alternatives:

3. You have the resources to look at more words
➔ Great! This can give you a much more reliable estimate

using k words instead of 1!

Def: $P(E_i)$ is Prob that word i occurs in an email (etc.)

Assume all $P(E_i)$ are independent of each other (etc.)

$P(S|w_1,w_2,\ldots,w_k) =$

$\prod_{i=1..k} P(E_i|S) * P(S) \; / \; \prod_{i=1..k} P(E_i) =$

$(\prod_{i=1..k} P(E_i|S * P(S)) + \prod_{i=1..k} P(E_i|\text{not-}S)*P(\text{not-}S))$

(As in the Tutorial, this may be simplified

if further assumptions are made)

Task: estimate P(S|E)

$P(S|E) = P(E|S) * P(S) / P(E)$

Other alternatives:

- synonyms: if "friend" indicates spam, then maybe "soul mate" too?
- sequences of n words
  (Rosen: compare "enhance performance" vs "operatic performance")
- take into account *how often* a word occurs
- etc.

Spam filtering

- Another example of Bayesian reasoning
  - » used for constructing a classifier (much like mushroom classification in the Lectures)
- Spammers second-guess spam filters, by adding "good" text to their messages (harvested from real non-spam messages, newspapers, etc.)
- Spam filters second-guess spammers doing this …
- A weapons race!