Learning from Data: Naive Bayes

Amos Storkey, School of Informatics

October 13, 2005

http://www.anc.ed.ac.uk/~amos/lfd/

Naive Bayes

- Typical example: “Bayesian Spam Filter”.
- Naive means naive. Bayesian methods can be much more sophisticated.
- Basic assumption: conditional independence.
- Given the class (eg “Spam”, “Ham”), whether one data item (eg word) appears is independent of whether another appears.
- Invariably wrong! But useful anyway.

Why?

- Easy to program. Simple and transparent.
- Fast to train. Fast to use.
- Can deal with uncertainty.
- Probabilistic.

Data types

- Naive Bayes assumption can use both continuous and discrete data.
- However generally understood in terms of discrete data.
- Binary and discrete very common. Do not use “1 of M”!
- E.g. Bag of words assumption for text classification:
- Can even mix different types of data

Bag of Words

- Each document is represented by a large vector.
- Each element of the vector represents the presence (1) or absence (0) of a particular word in the document.
- Certain words are more common in one document type than another.
- Can build another form of class conditional model using the conditional probability of seeing each word, given the document class (e.g. ham/spam).

Conditional Independence

- $P(X, Y) = P(X)P(Y|X)$.
- $P(X, Y|C) = P(X|C)P(Y|X, C)$. Think of C as a class label.
- The above is always true. However we can make an assumption
- $P(Y|X, C) = P(Y|C)$.
- Knowing about the value of X makes no difference to the value Y takes so long as we know the class C.
- We say that X and Y are conditionally independent given C.

Example

- Probability of a person hitting Jim (J) and a person hitting Michael (M) is most likely not independent.
- But they might be independent given that the person in question is (or is not) a known member of the class of bullies (B).
- $P(J, M) \neq P(J)P(M)$
- $P(J, M|B) = P(J|B)P(M|B)$.
- B explains all of the dependence between J and M.

Generally

- $x_1, x_2, \ldots, x_n$ are said to be conditionally independent given c iff
$$P(\mathbf{x}|c) = \prod_{i=1}^{n} P(x_i|c)$$
for $\mathbf{x} = (x_1, x_2, , \ldots, x_n)$.
- For example. We could have not just Jim and Michael, but Bob, Richard and Tim too.

Naive Bayes

- The equation on the previous slide is in fact the Naive Bayes Model.
$$P(\mathbf{x}|c) = \prod_{i=1}^{n} P(x_i|c)$$
for $\mathbf{x} = (x_1, x_2, , \ldots, x_n)$.
- The $\mathbf{x}$ is our attribute vector. And the c is our class label.
- We want to learn $P(c)$ and $P(x_i|c)$ from the data.
- We then want to find the best choice of c corresponding to a new datum (inference)
- The form of $P(x_i|c)$ is usually given. But we do need to learn the parameter.

Working Example

- See sheet section 3.
- Have a set of attributes.
- Inference first: Bayes rule.
- Learning the model $P(E)$,$P(S)$,$P(x|S)$,$P(x|E)$
- Naive Bayes assumption.

Problems with Naive Bayes

- 1 of M encoding
- Failed conditional independence assumptions.
- Worst case: repeated attribute.
- Double counted, triple counted etc.
- Conditionally dependent attributes can have too much influence.

Spam Example

- Bag of words.
- Probability of ham containing each word. Probability of spam containing each word.
- Prior probability of ham/spam.
- New document. Check the presence/absence of each word.
- Calculate the spam probability given the vector of word occurrence.
- How best to fool Naive Bayes? Introduce lots of hammy words into the document. Each hammy word is viewed independently and so they repeatedly count towards the ham probability.

Summary

- Conditional Independence
- Bag of Words
- Naive Bayes
- Learning Parameters
- Bayes Rule
- Working Examples