

Classification and Maximum Likelihood Learning

Noel Welsh

12 October 2010

1 Announcements

- The assignments are now up on the website. Your first assignment is due on the 25th of October. The task, essentially, is to conduct an experiment using one or more sensors in your robotics kit.
- Lab books. Get one.
- Masters (level 4) students have to do a research project. Check the web page.

2 Recap

Last lecture we looked at hypothesis testing. We started by formulating a hypothesis – that we had a fair coin. We then performed an experiment – twenty coin tosses, observing 12 heads and 8 tails. We then ran into the issue of analysing this results – does it support the hypothesis or not?

To answer this question we developed a probability model that explains the data – the binomial distribution. From this we could calculate the probability of the data given our hypothesis. This is known as the *data likelihood*, or just likelihood, written $P(\text{data}|H_0)$, where H_0 is the *null* hypothesis, in our case that the coin is fair.

We saw that the likelihood is the key quantity for testing a hypothesis. If the likelihood is below a certain threshold, called the *p-value*, then we reject the hypothesis. A common p-value is 0.05. For the experiment above the likelihood is 0.12, so we do not reject the hypothesis that we have a fair coin when using a p-value of 0.05.

We then saw that the binomial distribution converges to the normal distribution as the number of samples grows. This is a general property of the average of random variables that is formalised in the Central Limit Theorem.

3 Maximum Likelihood Learning

In summary, hypothesis testing is about choosing to accept or reject a given hypothesis based on how well it explains the problem. Today we're going to look at a related problem, that of finding the best hypothesis to explain some data.

Let's go back to the coin flip experiment. Remember that p , the probability of heads, is the important parameter (along with n , the number of flips) defining the binomial distribution. If you had to choose one value of p given our observations what would that value be? Why?

5 min group work.

Discuss. Key point: We can use the likelihood, $P(data|H)$, to choose between hypotheses, selecting that hypothesis that maximising the likelihood. This is called maximum likelihood learning.

4 Generative vs Discriminative Classification

For the main project, if you choose the task we have developed you'll want to distinguish red and blue cans. Imagine you've built a colour sensor (you can easily do this with three light sensors and coloured cellophane). How would you turn the output of your colour sensor into a colour classifier? What class of hypotheses does your classifier (implicitly) consider? Which hypothesis does it choose from that class?

5 min group work.

Get two teams up. Try to get a discriminative and a generative. Discuss the hypothesis classes as appropriate.

- In general there are two approaches to classification: generative and discriminative classifiers. Generative classifiers attempt to model all the data, while discriminative classifiers only try to model the difference between the classes.
- Formally, the problem is predicting class membership given data $P(c|data)$. Add in our restricted class of hypotheses, and we get $P(c|data, h)$.
- The task is to find $h \in H$ that maximises $P(c|data, h)$.
- The generative solution is to model the joint distribution $P(c, data|h)$ and then derive the conditional by application of *Bayes Rule*

- Bayes Rule is a general statement about probability

$$P(A|B) = \frac{P(A, B)}{P(B)}$$

$$= \frac{P(B|A)P(A)}{P(B)}$$

- From Bayes Rule we can derive the conditional:

$$P(c|data, h) = \frac{P(c, data|h)}{P(data|h)}$$

- The discriminative solution just models the conditional directly

4.1 Tradeoffs

Both generative and discriminative learning have advantages and disadvantages.

Can you guess some of them?

5 min group work.

My list:

4.1.1 Generative Models

- Advantages
 - Much more flexible model. Can form classifier, predictor, regressor from joint distribution. Can adapt to missing data, or new information. Example of the second: if we know there are fewer red cans, because we have collected them, we can reweight our predictions.
 - Elegant theory, though you haven't seen any of it (and probably won't).
 - Easily interpretable models, that give understanding of data.
 - Generally easier to train as we only have to model a class at a time.
- Disadvantages
 - Generally worse performance – we're not modelling the classification problem directly so there are more places to introduce error.

4.1.2 Discriminative Models

- Invert the above

5 Further Reading

- Tony Jebara's thesis "Discriminative, Generative and Imitative Learning" contains a modern overview of discriminative and generative learning. It is a bit different (and more advanced) than what I've presented. In particular what I call discriminative learning he calls conditional learning, and uses the term discriminative learning to refer to techniques that do away completely with notion of probabilities. This more nuanced view (probably) won't make it into this course.