6.858/18.428 Machine Learning **Lecture 11 : October 19th, 1994**

Lecturer: Ron Rivest *Scribe: Matthew Andrews*

11.1 Outline

- Bayesian learning
- MDL (Minimum Description Length)

In this lecture we consider the technique of Bayesian Inference. These notes are largely taken from Chapter 5 of *Lecture Notes in Machine Learning* by Ron Rivest.

11.2 Bayesian Inference – Introduction

In this chapter we review Bayesian inference. In many ways the theory of Bayesian inference is more advanced and sophisticated than the formal theories of learning that have developed within the computer science community; in other ways it is more primitive. I believe that it is instructive to include Bayesian inference in a study of machine learning in order to highlight these differences. Cheeseman has several interesting articles (e.g., [1, 2]) that make a good starting point for reviewing the issues. The journal containing [2] has many other articles discussing Bayesian approaches versus others.

Bayesian inference handles the following issues reasonably well:

1. Prior knowledge of the learner; these are reflected in the learner's set of *prior* probabilities.
2. Evidence that does not eliminate hypotheses, but which nonetheless gives preferential support to some hypotheses over others.
3. Degrees of belief other than certainty.
4. Consideration of how the learned information will be used.

Issues of computational complexity are usually not addressed within this framework.

Bayesian inference is well-suited to situations where the amount of data is sparse relative to the complexity of the hypothesis space, so that the given data does not uniquely specify the correct answer. When many possible answers remain, one needs a basis for deciding what to do. Bayesian inference is just such a theory.

In its simplest form, Bayesian inference is just the straightforward application of Bayes' Rule. Suppose we have a hypothesis space $\mathcal{H} = \{H_1, H_2, \ldots\}$ which consists of a set of hypotheses which are mutually exclusive (at most one of them is right) and exhaustive (at least one of them is right).

The learner associates with each hypothesis H_i a *prior probability* $P(H_i)$ according to the learner's background knowledge about the hypothesis space and the learning situation. These probabilities sum to one. Here $P(H_i)$ is the learner's initial *degree of belief* in the proposition that H_i is correct.

The learning situation then presents the learner with some data D, which provides some information about which hypothesis is correct. In our previous chapters, we only considered cases where D was either consistent with (or even determined by) each specific hypothesis, or inconsistent with each specific hypothesis, so that hypotheses were either eliminated or considered still viable. Here we refine this idea by introducing the *conditional probability* $P(D \mid H_i)$ that the data D would be experienced in the learning situation, *given that* H_i *is the correct hypothesis*. We assume that the learner can in principle compute $P(D \mid H_i)$ from D and H_i. The extreme cases of $P(D \mid H_i) = 1$ and $P(D \mid H_i) = 0$ correspond to our previous case where the data is either consistent with (forced by) the hypothesis or inconsistent with the hypothesis.

We can then define $P(D)$ to be the *unconditional* probability that the learner expects to see the data D:

$$P(D) = \sum_i P(D \mid H_i)P(H_i). \tag{11.1}$$

Bayes' Rule is just a straightforward fact about conditional probability:

$$P(H_i \mid D) = \frac{P(D \mid H_i)P(H_i)}{P(D)} \tag{11.2}$$

The interpretation of this rule is most interesting; it says that *the learner's final degree of belief in* H_i *is proportional to the product of his initial degree of belief in* H_i *and the conditional probability of seeing* D *given* H_i. (The denominator $P(D)$ can be considered as just a normalizing constant, so that the final probabilities add up to one.)

What should the learner do if there is more than one hypothesis consistent with the observed data D, and he is asked to specify some single hypothesis as his "inference"? There are two standard answers to this question:

- *(ML – Maximum Likelihood)* Choose the hypothesis which has maximum likelihood $P(D \mid H_i)$ given the data D.
- *(MAP – Maximum A Posteriori)* Choose the hypothesis which has maximum *a posteriori* (i.e. final) probability $P(H_i \mid D)$.

The MAP method is the true Bayesian approach, since the ML technique doesn't even make use of Bayes' Rule. The second approach is better (in my opinion), but the first is often used when one doesn't want to work with prior probabilities for some reason. The approaches coincide when the prior distribution weights each hypothesis equally.

If the learner is asked to specify a particular hypothesis, and there is a known cost for being wrong (i.e. the cost of outputting H_i when H_j is the truth), one can use the known set of final probabilities to minimize the expected cost of making an error.

If the learner is not asked to specify a particular hypothesis as his choice, but is instead asked to make a prediction based on his current (final) set of beliefs, he can apply *transduction* (see Cheeseman [10]). Using $P(H_i)$ now to denote his current (as updated) set of probabilities, this is merely the use of equation (11.1) to predict the probability of any proposed new data D. For example, if we are inferring Boolean functions, and there are many more-or-less equally likely functions consistent with the data which predict different values at a new point x, then we are not going to get a strong prediction for the value of the unknown function at x.

11.3 A Coin Example

Suppose that we have a coin and we know that it is either fair or else it has a 60% bias in favour of heads. That is, our two hypotheses are:

- H_1 - fair coin
- H_2 - 60% heads, 40% tails.

Usually when we are given a coin we assume that it is more likely to be fair than to have a 60% bias. Hence we take for our *a priori* probabilities $P(H_1) = 0.75$,

$P(H_2) = 0.25$. Suppose that our first piece of evidence D_1 is a single coin toss which comes up heads.

$$\begin{array}{rclrcl} P(D_1|H_1) & = & 0.5 & P(D_1|H_2) & = & 0.60 \\ P(H_1)P(D_1|H_1) & = & 0.375 & P(H_2)P(D_1|H_2) & = & 0.15 \\ \Rightarrow P(H_1|D_1) & = & 0.714 & P(H_2|D_1) & = & 0.286. \end{array}$$

Hence after seeing one head we still believe that a fair coin is the most likely hypothesis. Our second piece of evidence is a sequence of 100 coins tosses which contains 70 heads. Note that we have seen a specific sequence of this form and so $P(D_2|H_1) = 2^{-100}$ (not $\binom{100}{70}2^{-100}$). Similarly, $P(D_2|H_2) = (0.6)^{70}(0.4)^{30}$. i.e.,

$$\begin{array}{rclrcl} P(D_2|H_1) & = & 7.9 \times 10^{-31} & P(D_2|H_2) & = & 3.41 \times 10^{-28} \\ P(H_1|D_1)P(D_2|H_1) & = & 5.63 \times 10^{-31} & P(H_2|D_1)P(D_2|H_2) & = & 9.75 \times 10^{-29} \\ \Rightarrow P(H_1|D_1, D_2) & = & 0.0057 & P(H_2|D_1, D_2) & = & 0.9943. \end{array}$$

If our goal is to output a hypothesis then after the first flip we would predict H_1. After the 101st flip we would predict H_2. If however our goal is to make predictions then after the 101st flip we would believe that the probability of heads is $(0.0057)(0.5) + (0.9943)(0.6) > 0.5$ and so we would predict heads.

11.4 Probabilities as Degrees of Belief

What is a *probability*?

There are a number of different kinds of probabilities, such as "physical probability", "long run chance of success", or "subjective probability". As probability theory has matured, it has become recognized that the most fundamental notion is that of *subjective probability*, which represents a *degree of belief*. The other notions can be viewed as derivative of this primary notion. See Good[3] for an excellent discussion of this issue.

The reason for mentioning this fact is that some people feel squeamish about using probabilities to measure degrees of belief, since elementary courses in probability theory often work with the maximum-likelihood approach (and so avoid the use of prior or final probabilities), or with situations where the prior probabilities are given externally somehow. It is not often emphasized that the notion of degree of belief is really at the heart of probability theory. See Cheeseman [1] for a fuller discussion of the role of probability theory in AI.

11.5 Prior Probabilities

Prior probabilities are mildly controversial, in the sense that the theory doesn't say much about how they are to be created, merely how they are to be updated using Bayes' Rule. This resultant ambiguity motivates some people to avoid the issue altogether and stick to a Maximum Likelihood approach.

Jaynes [5, 6] gives an nice overview of the problem, together with an introduction to the principle of "Maximum Entropy" as a means for constructing prior probability distributions. The principle of maximum entropy suggests that one should use as a prior the distribution with maximum entropy (i.e., choose priors to maximise $\sum_i P(H_i)\log(\frac{1}{P(H_i)})$) subject to any known constraints. This is a formal way of picking the distribution which expresses maximum ignorance or uncertainty, subject to the given constraints.

As an example of an application of the Maximum Entropy Principle, suppose that you have a (six-sided) die, where you know that the die is unfair since you know that the average roll is 4.5. That is, if p_i is the chance of rolling i, we know that $\sum_{i=1}^{6} ip_i = 4.5$. However, we know nothing else about the die. *What is the chance of rolling a 6?*

The problem is clearly underspecified, and there is no "right" answer. We can construct a "maximally ignorant" prior probability distribution by finding the distribution which maximizes $\sum_{i=1}^{6} -p_i \log(p_i)$ subject to the given constraint:

$$(p_1, \ldots, p_6) = (0.0543, 0.0788, 0.1142, 0.1654, 0.2398, 0.3475). \tag{11.3}$$

The answer to our question is thus: *we expect to see a 6 with probability 0.3475.*

In a similar spirit, Rissanen [9] suggests using the following prior probability distribution if the hypothesis space $\mathcal{H} = \{H_1, H_2, \ldots,\}$ is countably infinite:

$$P(H_i) = \frac{1}{cf^*(i)}; \tag{11.4}$$

here $f^*(n)$ is defined by the equation

$$f^*(n) = n \cdot \log n \cdot \log\log n \cdots, \tag{11.5}$$

where only the positive terms are included in the product, and $c = 2.865064\ldots$ is a normalizing constant. This probability distribution goes to 0 as "slowly as possible" subject to the condition that $\sum_i P(H_i) = 1$, and so is thus another way of assuming "maximum ignorance".

11.6 Prior Probabilities and Simplicity

From another viewpoint, we can view the prior probability distribution as a mechanism for expressing opinions about the *simplicity* of hypotheses. Suppose we have a situation where the each conditional probability $P(D \mid H_i)$ is either 0 or 1 — the data either disqualifies a hypothesis or is determined by the hypothesis. In this case, in the absence of sufficient information to select a unique hypothesis, one would like to select the "simplest" hypothesis which is consistent with the data. The MAP principle will select the one with maximum prior probability. Thus we can use the prior probability distribution as a means to differentiate hypotheses according to their simplicity, with simpler hypthoses getting larger prior probabilities. For example, with Rissanen's prior, hypothesis H_i is given a higher initial probability that H_j if $i < j$, so the ordering $\{H_1, H_2, \ldots\}$ could be interpreted as a listing of the hypothesis space in order of decreasing simplicity (increasing complexity).

11.7 Minimum Description Length (MDL)

Bayes Rule can be reformulated into an additive form using logarithms: let us call

$$L(H_i) = -\log(P(H_i)) \tag{11.6}$$

the initial (or prior) *complexity* of hypothesis H_i, in line with our discussion above. (This term is nonstandard.)

Similarly, let us call

$$L(D \mid H_i) = -\log(P(D \mid H_i)) \tag{11.7}$$

the *weight of evidence against* H_i provided by D. Using base-two logarithms, we can measure both complexities and weights of evidence in *bits*.

If we define the *final complexity* of H_i, given D, as

$$L(H_i \mid D) = -\log(P(H_i \mid D)) \tag{11.8}$$

then Bayes' Rule can be reinterpreted in additive form:

$$L(H_i \mid D) = L(H_i) + L(D \mid H_i) + c \tag{11.9}$$

where c is a suitable normalizing constant (which is a function of D and the priors only).

In some applications it is convenient to work with unnormalized complexities $L'(H_i)$ only, since the relative probabilities of various hypotheses depend only on the difference between their relative complexities.

In any case, we see that Bayes' Rule provides a tradeoff between the complexity of the hypothesis and the "fit" of the data to that hypothesis.

11.8 A Virus Example

Suppose a person believes that he has a prior probability of carrying the AIDS virus of 10^{-5}, due to having had a blood transfusion before blood screening was routine. Suppose that a blood test is 99% accurate for carrying the virus, and that our poor friend tests positive. (This is the data D.) How likely is it that he actually has the virus? Using base 10 logarithms, we have

$$\begin{aligned}
L(\text{virus}) &= -\log(10^{-5}) = 5 \\
L(\overline{\text{virus}}) &= -\log(0.99999) = 4.3 \times 10^{-6} \\
L(D \mid \text{virus}) &= \log(0.99) = 4.3 \times 10^{-3} \\
L(D \mid \overline{\text{virus}}) &= \log(0.01) = 2 \\
L'(\text{virus} \mid D) &= 5 + 4.3 \times 10^{-3} = 5.043 \\
L'(\overline{\text{virus}} \mid D) &= 4.3 \times 10^{-6} + 2 = 2.000043
\end{aligned}$$

Thus, even after the positive test, our friend still is 1000 times more likely *not* to have the virus than to have it, since $1000 \approx 10^{(5.043-2.000043)}$.

(This sort of analysis is obviously relevant when one begins to consider the proposal to institute wide-spread testing for AIDS.)

11.9 The coin example revisited

Recall that H_1 was the hypothesis that we have a fair coin. H_2 was the hypothesis that the coin has a 60% bias in favour of heads. D_1 was the first coin toss which came up heads. D_2 was a sequence of 100 coin tosses which contained 70 heads. The initial probabilities were $P(H_1) = 0.75$, $P(H_2) = 0.25$. We now have,

$$\begin{array}{rcccrcc}
L(H_1) & = & 0.415 & & L(H_2) & = & 2 \\
L(D_1|H_1) & = & 1 & & L(D_1|H_2) & = & 0.737 \\
\Rightarrow L(H_1) + L(D_1|H_1) & = & 1.415 & & L(H_2) + L(D_1|H_2) & = & 2.737 \\
\\
L(D_2|H_2) & = & 100 & & L(D_2|H_2) & = & 91.244 \\
\Rightarrow L(H_1|D_1) + L(D_2|H_2) & = & 101.415 & & L(H_2|D_1) + L(D_2|H_2) & = & 93.981
\end{array}$$

11.10 Relating MDL and MAP

One can reverse the relationship between prior probabilities and complexity. That is, rather than determine the complexity of the hypotheses from the prior probabilities, one can go in the other direction and use a measure of the complexity of the hypotheses to determine the prior probabilities, using equation (11.6).

For example, Rissanen [8, 9] proposes that one can use *syntactic* measures of the complexity of hypothesis – how many bits does it take to encode the hypothesis? This approach is very natural for a computer scientist, since devising encodings for hypotheses seems easier to think about than devising prior probabilities. This approach has been examined further by Hart [4].

Similarly, Rissanen proposes determining the weight of evidence against H_i provided by D by actually encoding D assuming H_i as given.

Finally, the combined "description length" for D using H_i is just the length of the encoding of H_i plus the length of D using H_i as given. Except for the normalization factor (which is fixed once D is known), this is seen to be the same as the final complexity as determined by equation (11.9).

The hypothesis which gives D the minimum description length (i.e. which "compresses" D the most, when the cost of encoding the hypothesis is counted as well) is chosen as the "best" hypothesis according to the Minimum Description Length Principle.

We see that we can view Rissanen's Minimum Description Length Principle as a special case of Bayes' Rule.

Furthermore, we see that Bayesian inference and data compression are related: a good theory about the data allows it to be compressed better.

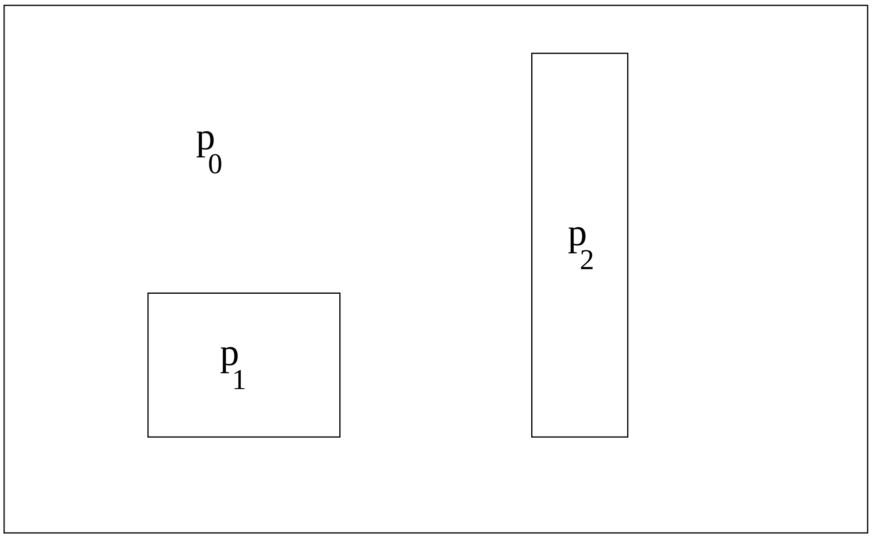

Figure 11.1: A hypothesis with three probabilities.

11.11 A Cosmology Example

Suppose that we would like to know whether or not galaxies are evenly spread about the universe. We first of all divide the universe into an $n \times n$ grid. We have the following hypotheses.

- H_{p_0} is the hypothesis that there is no structure, every cell has a galaxy with probability p_0.
- $H_{p_0,((a_1,a_2),(b_1,b_2),p_1)}$ is the hypothesis that in the rectangle with lower left corner (a_1, a_2) and upper right corner (b_1, b_2) there is a galaxy with probability p_1. Elsewhere there is a galaxy with probability p_0.
- $H_{p_0,\left\{\left((a_1^{(i)},a_2^{(i)}),(b_1^{(i)},b_2^{(i)}),p_i\right)\right\}}$ is the hypothesis that in the rectangle with lower left corner $(a_1^{(i)}, a_2^{(i)})$ and upper right corner $(b_1^{(i)}, b_2^{(i)})$ there is a galaxy with probability p_i. A cell which is outside all of the rectangles has a galaxy with probability p_0. Here we assume that $i \in \{1, 2, \ldots, k\}$ and the rectangles are disjoint.

Let H be the entropy function. That is, $H(p) = p\log(\frac{1}{p}) + (1-p)\log(\frac{1}{1-p})$. Let λ be the number of bits that we are using to encode probabilities and let D be the distribution of galaxies. The following formulas give the expected combined description lengths for D for each of the hypotheses.

- If the hypothesis is H_{p_0} then the expected length is $\lambda + n^2 H(p_0)$. λ bits are used to represent p_0 and $n^2 H(p_0)$ is the expected number of bits required to represent D.
- If the hypothesis is $H_{p_0,\left\{\left((a_1^{(i)},a_2^{(i)}),(b_1^{(i)},b_2^{(i)}),p_i\right)\right\}}$ then the expected length is,

  $$(k+1)\lambda + 4k\log n + \sum A_i H(p_i) + (n^2 - \sum A_i)H(p_0),$$

  where A_i is the area of the ith rectangle. The first term in the above formula is the number of bits needed to represent the probabilities. The second term corresponds to representing the rectangles. The last two terms are the expected number of bits needed to represent D.

If we wanted to give prior probabilities to the hypotheses then we would let,

$$\begin{aligned} P(H_{p_0}) &= c2^{-\lambda}, \\ P\left(H_{p_0,\left\{\left((a_1^{(i)},a_2^{(i)}),(b_1^{(i)},b_2^{(i)}),p_i\right)\right\}}\right) &= c2^{-(k+1)\lambda-4k\log n}, \end{aligned}$$

where c is a suitable normalizing constant.

11.12 Summary

Bayesian inference provides a way of handling the uncertainty that arises from hypotheses of differing initial complexity and from uneven conditional probabilities. It also provides a way of formulating the tradeoff between the complexity of a hypothesis and its fit to the data.

11.13 Bibliography

[1] Peter C. Cheeseman. In defense of probability. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pages 1002–1009, 1985.

[2] Peter C. Cheeseman. An inquiry into computer understanding. *Computational Intelligence*, 4(1):58–66, February 1988.

[3] I. J. Good. Kinds of probability. *Science*, 129(3347):443–447, February 1959.

[4] George W. Hart. *Minimum Information Estimation of Structure.* PhD thesis, MIT Dept. of Electrical Engineering and Computer Science, April 1987. Appears as LIDS-TH-1664.

[5] Edwin T. Jaynes. Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics*, SSC-4(3):227–241, September 1968.

[6] Edwin T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9):939–952, September 1982.

[7] A. Levin, Leonid. A concept of independence with applications in various fields of mathematics. Technical Report MIT/LCS/TR-235, MIT Laboratory for Computer Science, April 1980.

[8] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[9] Jorma Rissanen. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 11(2):416–431, 1983.

[10] Matthew Self and Peter C. Cheeseman. Bayesian prediction for artificial intelligence. (unpublished manuscript).