

# Artificial Intelligence Programming

## Bayesian Learning

Chris Brooks

Department of Computer Science  
University of San Francisco

## Learning and Classification

- An important sort of learning problem is the *classification* problem.
- This involves placing examples into one of two or more classes.
  - Should/shouldn't get credit
  - Categories of documents.
  - Golf-playing/non-golf-playing days
- This requires access to a set of *labeled* training examples, which allow us to induce a hypothesis that describes how to decide what class an example should be in.

## Bayes' Theorem

- Recall the definition of Bayes' Theorem
- $P(b|a) = \frac{P(a|b)P(b)}{P(a)}$
- Let's rewrite this a bit.
- Let  $D$  be the data we've seen so far.
- Let  $h$  be a possible hypothesis
- $P(h|D) = \frac{P(D|h)P(h)}{P(D)}$

Department of Computer Science — University of San Francisco — p.1/77

Department of Computer Science — University of San Francisco — p.1/77

## MAP Hypothesis

- Usually, we won't be so interested in the particular probabilities for each hypothesis.
- Instead, we want to know: Which hypothesis is most likely, given the data?
  - Which classification is the most probable?
  - Is *PlayTennis* or  $\neg$ *PlayTennis* more likely?
- We call this the *maximum a posteriori hypothesis* (MAP hypothesis).
- In this case, we can ignore the denominator in Bayes' Theorem, since it will be the same for all  $h$ .
- $h_{MAP} = \operatorname{argmax}_{h \in H} P(D|h)P(h)$

Department of Computer Science — University of San Francisco — p.3/77

## MAP Hypothesis

- Advantages:
  - Simpler calculation
  - No need to have a prior for  $P(D)$

Department of Computer Science — University of San Francisco — p.4/77

## ML Hypothesis

- In some cases, we can simplify things even further.
- What are the priors  $P(h)$  for each hypothesis?
- Without any other information, we'll often assume that they're equally possible.
  - Each has probability  $\frac{1}{H}$
- In this case, we can just consider the conditional probability  $P(D|h)$ .
- We call the hypothesis that maximizes this conditional probability the *maximum likelihood hypothesis*.
- $h_{ML} = \operatorname{argmax}_{h \in H} P(D|h)$

Department of Computer Science — University of San Francisco — p.5/77

## Example

- Imagine that we have a large bag of candy. We want to know the ratio of cherry to lime in the bag.
- We start with 5 hypotheses:
  - $h_1$ : 100% cherry
  - $h_2$ : 75% cherry, 25% lime.
  - $h_3$ : 50% cherry, 50% lime
  - $h_4$ : 25% cherry, 75% lime
  - $h_5$ : 100% lime
- Our agent repeatedly draws pieces of candy.
- We want it to correctly pick the type of the next piece of candy.

## Example

- Let's assume our priors for the different hypotheses are:
  - $(0.1, 0.2, 0.4, 0.2, 0.1)$
- Also, we assume that the observations are i.i.d.
  - This means that each choice is independent of the others. (order doesn't matter)
- In that case, we can multiply probabilities.
- $P(D|h_i) = \prod_j P(d_j|h_i)$
- Suppose we draw 10 limes in a row.  $P(D|h_3)$  is  $(\frac{1}{2})^{10}$ , since the probability of drawing a lime under  $h_3$  is  $\frac{1}{2}$ .

## Example

- How do the hypotheses change as data is observed?
- Initially, we start with the priors:  $(0.1, 0.2, 0.4, 0.2, 0.1)$
- Then we draw a lime.
  - $P(h_1|lime) = \alpha P(lime|h_1)P(h_1) = 0$ .
  - $P(h_2|lime) = \alpha P(lime|h_2)P(h_2) = \alpha \frac{1}{4} * 0.2 = \alpha 0.05$ .
  - $P(h_3|lime) = \alpha P(lime|h_3)P(h_3) = \alpha \frac{1}{2} * 0.4 = \alpha 0.2$
  - $P(h_4|lime) = \alpha P(lime|h_4)P(h_4) = \alpha \frac{3}{4} * 0.2 = \alpha 0.15$ .
  - $P(h_5|lime) = \alpha P(lime|h_5)P(h_5) = \alpha 1 * 0.1 = \alpha 0.1$ .
  - $\alpha = 2$ .

## Example

- Then we draw a second lime.
  - $P(h_1|lime, lime) = \alpha P(lime, lime|h_1)P(h_1) = 0$ .
  - $P(h_2|lime, lime) = \alpha P(lime, lime|h_2)P(h_2) = \alpha \frac{1}{4} * 0.2 = \alpha 0.0125$ .
  - $P(h_3|lime, lime) = \alpha P(lime, lime|h_3)P(h_3) = \alpha \frac{1}{2} * 0.4 = \alpha 0.1$
  - $P(h_4|lime, lime) = \alpha P(lime, lime|h_4)P(h_4) = \alpha \frac{3}{4} * 0.2 = \alpha 0.1125$ .
  - $P(h_5|lime) = \alpha P(lime|h_5)P(h_5) = \alpha 1 * 0.1 = \alpha 0.1$ .
  - $\alpha = 3.07$ .
- Strictly speaking, we don't really care what  $\alpha$  is.
- We can just select the MAP hypothesis, since we just want to know the most likely hypothesis.

## Bayesian Learning

- Eventually, the true hypothesis will dominate all others.
  - Caveat: assuming the data is noise-free, or noise is uniformly distributed.
- Notice that we can use Bayesian learning (in this case) either as a batch algorithm or as an incremental algorithm.
- We can always easily update our hypotheses to incorporate new evidence.
  - This depends on the assumption that our observations are independent.

## Learning bias

- What sort of bias does Bayesian Learning use?
- Typically, simpler hypotheses will have larger priors.
- More complex hypotheses will fit data more exactly (but there's many more of them).
  - Under these assumptions,  $h_{MAP}$  will be the simplest hypothesis that fits the data.
  - This is Occam's razor, again.
  - Think about the deterministic case, where  $P(h_i|D)$  is either 1 or 0.

## Bayesian Concept Learning

- Bayesian Learning involves estimating the likelihood of each hypothesis.
- In a more complex world where observations are not independent, this could be difficult.
- Our first cut at doing this might be a brute force approach:
  1. For each  $h$  in  $H$ , calculate  $P(h|D) = \frac{P(D|h)P(h)}{P(D)}$
  2. From this, output the hypothesis  $h_{MAP}$  with the highest posterior probability.
- This is what we did in the example.
  - Challenge - Bayes' Theorem can be computationally expensive to use when observations are not i.i.d.

- $P(h|o_1, o_2) = \frac{P(o_1|h, o_2)P(h|o_2)}{P(o_1|o_2)}$

Department of Computer Science — University of San Francisco — p.12/77

## Bayesian Optimal Classifiers

- There's one other problem with the formulation as we have it.
- Usually, we're not so interested in the hypothesis that fits the data.
- Instead, we want to classify some unseen data, given the data we've seen so far.
- One approach would be to just return the MAP hypothesis.
- We can do better, though.

Department of Computer Science — University of San Francisco — p.13/77

## Bayesian Optimal Classifiers

- Suppose we have three hypotheses and posteriors:  
 $h_1 = 0.4, h_2 = 0.3, h_3 = 0.3$ .
- We get a new piece of data -  $h_1$  says it's positive,  $h_2$  and  $h_3$  negative.
- $h_1$  is the MAP hypothesis, yet there's a 0.6 chance that the data is negative.
- By combining weighted hypotheses, we improve our performance.

Department of Computer Science — University of San Francisco — p.14/77

## Bayesian Optimal Classifiers

- By combining the predictions of each hypothesis, we get a Bayesian optimal classifier.
- More formally, let's say our unseen data belongs to one of  $v$  classes.
- The probability  $P(v_j|D)$  that our new instance belongs to class  $v_j$  is:
  - $\sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$
- Intuitively, each hypothesis gives its prediction, weighted by the likelihood that that hypothesis is the correct one.
- This classification method is provably optimal - on average, no other algorithm can perform better.

Department of Computer Science — University of San Francisco — p.15/77

## Problems with the Bayes Optimal Classifier

- However, the Bayes optimal classifier is mostly interesting as a theoretical benchmark.
- In practice, computing the posterior probabilities is exponentially hard.
- This problem arises when instances or data are conditionally dependent upon each other.
- Can we get around this?

Department of Computer Science — University of San Francisco — p.16/77

## Naive Bayes classifier

- The Naive Bayes classifier makes a strong assumption that makes the algorithm practical:
  - Each attribute of an example is independent of the others.
  - $P(a \wedge b) = P(a)P(b)$  for all  $a$  and  $b$ .
- This makes it straightforward to compute posteriors.

Department of Computer Science — University of San Francisco — p.17/77

## The Bayesian Learning Problem

- Given: a set of labeled, multivalued examples.
- Find a function  $F(x)$  that correctly classifies an unseen example with attributes  $(a_1, a_2, \dots, a_n)$ .
- Call the most probable category  $v_{map}$ .
- $v_{map} = \operatorname{argmax}_{v_i \in V} P(v_i | a_1, a_2, \dots, a_n)$
- We rewrite this with Bayes' Theorem as:  

$$v_{map} = \operatorname{argmax}_{v_i \in V} P(a_1, a_2, \dots, a_n | v_i) P(v_i)$$
- Estimating  $P(v_i)$  is straightforward with a large training set; count the fraction of the set that are of class  $v_i$ .
- However, estimating  $P(a_1, a_2, \dots, a_n | v_i)$  is difficult unless our training set is *very* large. We need to see every possible attribute combination many times.

## Naive Bayes assumption

- Naive Bayes assumes that all attributes are conditionally independent of each other.
- In this case,  $P(a_1, a_2, \dots, a_n | v_i) = \prod_i P(a_i | v_i)$ .
- This can be estimated from the training data.
- The classifier then picks the class with the highest probability according to this equation.
- Interestingly, Naive Bayes performs well even in cases where the conditional independence assumption fails.

## Example

- Recall your tennis-playing problem from the decision tree homework.
- We want to use the training data and a Naive Bayes classifier to classify the following instance:
- Outlook = Sunny, Temperature = Cool, Humidity = high, Wind = Strong.

## Example

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

## Example

- Our priors are:
  - $P(\text{PlayTennis} = \text{yes}) = 9/14 = 0.64$
  - $P(\text{PlayTennis} = \text{no}) = 5/14 = 0.36$
- We can estimate:
  - $P(\text{wind} = \text{strong} | \text{PlayTennis} = \text{yes}) = 3/9 = 0.33$
  - $P(\text{wind} = \text{strong} | \text{PlayTennis} = \text{no}) = 3/5 = 0.6$
  - $P(\text{humidity} = \text{high} | \text{PlayTennis} = \text{yes}) = 3/9 = 0.33$
  - $P(\text{humidity} = \text{high} | \text{PlayTennis} = \text{no}) = 4/5 = 0.8$
  - $P(\text{outlook} = \text{sunny} | \text{PlayTennis} = \text{yes}) = 2/9 = 0.22$
  - $P(\text{outlook} = \text{sunny} | \text{PlayTennis} = \text{no}) = 3/5 = 0.6$
  - $P(\text{temp} = \text{cool} | \text{PlayTennis} = \text{yes}) = 3/9 = 0.33$
  - $P(\text{temp} = \text{cool} | \text{PlayTennis} = \text{no}) = 1/5 = 0.2$

## Example

- $v_{yes} = P(\text{yes})P(\text{sunny} | \text{yes})P(\text{cool} | \text{yes})P(\text{high} | \text{yes})P(\text{strong} | \text{yes}) = 0.005$
- $v_{no} = P(\text{no})P(\text{sunny} | \text{no})P(\text{cool} | \text{no})P(\text{high} | \text{no})P(\text{strong} | \text{no}) = 0.0206$
- So we see that not playing tennis is the maximum likelihood hypothesis.
- Further, by normalizing, we see that the classifier predicts a  $\frac{0.0206}{0.005+0.0206} = 0.80$  probability of not playing tennis.

## Estimating Probabilities

- As we can see from this example, estimating probabilities through frequency is risky when our data set is small.
- We only have 5 negative examples, so we may not have an accurate estimate.
- A better approach is to use the following formula, called an  $m$ -estimate:
  - $\frac{n_c + mp}{n + m}$
- Where  $n_c$  is the number of individuals with the characteristic of interest (say Wind = strong),  $n$  is the total number of positive/negative examples,  $p$  is our prior estimate, and  $m$  is a constant called the *equivalent sample size*.

## Estimating Probabilities

- $m$  determines how heavily to weight  $p$ .
- $p$  is assumed to be uniform.
- So, in the Tennis example,  
$$P(\text{wind} = \text{strong} | \text{playTennis} = \text{no}) = \frac{3 + 0.2m}{5 + m}$$
- We'll determine an  $m$  based on sample size.
  - If  $m$  is zero, we just use observed data.
  - If  $m \gg n$ , we use the prior.
  - Otherwise  $m$  lets us weight these parameters' relative influence.

## Using Naive Bayes to classify spam

- One area where Naive Bayes has been very successful is in text classification.
  - Despite the violation of independence assumptions.
- Classifying spam is just a special case of text classification.
- Problem - given some emails labeled ham or spam, determine the category of new and unseen documents.
- Our features will be the tokens that appear in a document.
- Based on this, we'll predict a category.

## Using Naive Bayes to classify spam

- For a given email, we'll want to compute the MAP hypothesis - that is, is:
  - $P(\text{spam} | t_1, t_2, \dots, t_n)$  greater than
  - $P(\text{ham} | t_1, t_2, \dots, t_n)$
- We can use Bayes' rule to rewrite these as:
  - $\alpha P(t_1, t_2, \dots, t_n | \text{spam}) P(\text{spam})$
  - $\alpha P(t_1, t_2, \dots, t_n | \text{ham}) P(\text{ham})$

## Using Naive Bayes to classify spam

- We can then use the Naive Bayes assumption to rewrite these as:
  - $\alpha P(t_1 | \text{spam}) P(t_2 | \text{spam}) \dots P(t_n | \text{spam}) P(\text{spam})$
  - $\alpha P(t_1 | \text{ham}) P(t_2 | \text{ham}) \dots P(t_n | \text{ham}) P(\text{ham})$
- And this we know how to compute.

## Using Naive Bayes to classify spam

- We can get the conditional probabilities by counting tokens in the training set.
- We can get the priors from the training set, or through estimation.

## Using Naive Bayes to classify spam

- There are a lot of wrinkles to consider:
  - What should be treated as a token? All words? All strings? Only some words?
  - Should headers be given different treatment? Greater or less emphasis?
  - What about HTML?
  - When classifying an email, should you consider all tokens, or just the most significant?
  - When computing conditional probabilities, should you could the fraction of documents a token appear in, or the fraction of words represented by a particular token?
- These are for you to decide ...