Bayesian Learning

Unit 3

Introduction

- Bayesian reasoning provides a probabilistic approach to inference.
- It is based on the assumption that the quantities of interest are governed by probability distributions and that optimal decisions can be made by reasoning about these probabilities together with observed data.
- Bayesian learning methods are relevant to our study of machine learning for two different reasons.
  - Bayesian learning algorithms that calculate explicit probabilities for hypotheses, such as the naive Bayes classifier, are among the most practical approaches to certain types of learning problems.
  - they provide a useful perspective for understanding many learning algorithms that do not explicitly manipulate probabilities

Features of Bayesian learning methods

- Each observed training example can incrementally decrease or increase the estimated probability that a hypothesis is correct.
- Prior knowledge can be combined with observed data to determine the final probability of a hypothesis.
- Bayesian methods can accommodate hypotheses that make probabilistic predictions
- New instances can be classified by combining the predictions of multiple hypotheses, weighted by their probabilities.
- Even in cases where Bayesian methods prove computationally intractable, they can provide a standard of optimal decision making against which other practical methods can be measured.

BAYES THEOREM

- A **prior probability** is an initial probability value originally obtained before any additional information is obtained.
- A **posterior probability** is a probability value that has been revised by using additional information that is later obtained.

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- $P(h)$ = prior probability of hypothesis h
- $P(D)$ = prior probability of training data D
- $P(h|D)$ = probability of h given D
- $P(D|h)$ = probability of D given h

Choosing Hypotheses

Generally want the most probable hypothesis given the training data

Maximum a posteriori hypothesis h_{MAP}:

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} P(h|D) \\ &= \arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\ &= \arg\max_{h \in H} P(D|h)P(h) \end{aligned}$$

If assume $P(h_i) = P(h_j)$ then can further simplify, and choose the *Maximum likelihood* (ML) hypothesis

$$h_{ML} = \arg\max_{h_i \in H} P(D|h_i)$$

Example

- We have prior knowledge that over the entire population of people only .008 have this disease. The test returns a correct positive result in only 98% of the cases in which the disease is actually present and a correct negative result in only 97% of the cases in which the disease is not present.

$$P(cancer) = .008, \quad P(\neg cancer) = .992$$
$$P(\oplus|cancer) = .98, \quad P(\ominus|cancer) = .02$$
$$P(\oplus|\neg cancer) = .03, \quad P(\ominus|\neg cancer) = .97$$

Suppose we now observe a new patient for whom the lab test returns a positive result. Should we diagnose the patient as having cancer or not?

$$P(\oplus|cancer)P(cancer) = (.98).008 = .0078$$
$$P(\oplus|\neg cancer)P(\neg cancer) = (.03).992 = .0298$$
$$\text{Thus, } h_{MAP} = \neg cancer.$$

Example

- In Orange County, 51% of the adults are males. (It doesn't take too much advanced mathematics to deduce that the other 49% are females.) One adult is randomly selected for a survey involving credit card usage.
- Find the prior probability that the selected person is a male.

- Let's use the following notation:
- M = male F = female (or not male)
- we know that 51% of the adults in Orange County are males, so the probability of randomly selecting an adult and getting a male is given by $P(M) = 0.51$.

Part B

- It is later learned that the selected survey subject was smoking a cigar. Also, 9.5% of males smoke cigars, whereas 1.7% of females smoke cigars (based on data from the Substance Abuse and Mental Health Services Administration). Use this additional information to find the probability that the selected subject is a male.

- Based on the additional given information, we have the following:
- $P(M) = 0.51$ because 51% of the adults are males
- $P(F) = 0.49$ because 49% of the adults are females (not males)
- $P(C|M) = 0.095$ because 9.5% of the males smoke cigars (That is, the probability of getting someone who smokes cigars, given that the person is a male, is 0.095.)
- $P(C|F) = 0.017$. because 1.7% of the females smoke cigars (That is, the probability of getting someone who smokes cigars, given that the person is a female, is 0.017.)

$$P(M \mid C) = \frac{P(M) \cdot P(C \mid M)}{[P(M) \cdot P(C \mid M)] + [P(\overline{M}) \cdot P(C \mid \overline{M})]}$$

$$= \frac{0.51 \cdot 0.095}{[0.51 \cdot 0.095] + [0.49 \cdot 0.017]}$$

$$= 0.85329341$$

$$= 0.853 \text{ (rounded)}$$

Example

An aircraft emergency locator transmitter (ELT) is a device designed to transmit a signal in the case of a crash. The Altigauge Manufacturing Company makes 80% of the ELTs, the Bryant Company makes 15% of them, and the Chartair Company makes the other 5%. The ELTs made by Altigauge have a 4% rate of defects, the Bryant ELTs have a 6% rate of defects, and the Chartair ELTs have a 9% rate of defects (which helps to explain why Chartair has the lowest market share).

a. If an ELT is randomly selected from the general population of all ELTs, find the probability that it was made by the Altigauge Manufacturing Company.

Solution

We use the following notation:

A = ELT manufactured by Altigauge
B = ELT manufactured by Bryant
C = ELT manufactured by Chartair

D = ELT is defective
$\overline{D}$ = ELT is not defective (or it is good)

If an ELT is randomly selected from the general population of all ELTs, the probability that it was made by Altigauge is 0.8 (because Altigauge manufactures 80% of them).

b. If a randomly selected ELT is then tested and is found to be defective, find the probability that it was made by the Altigauge Manufacturing Company.

$P(\text{A}) = 0.80$	because Altigauge makes 80% of the ELTs
$P(\text{B}) = 0.15$	because Bryant makes 15% of the ELTs
$P(\text{C}) = 0.05$	because Chartair makes 5% of the ELTs
$P(\text{D}\|\text{A}) = 0.04$	because 4% of the Altigauge ELTs are defective
$P(\text{D}\|\text{B}) = 0.06$	because 6% of the Bryant ELTs are defective
$P(\text{D}\|\text{C}) = 0.09$	because 9% of the Chartair ELTs are defective

Here is Bayes' theorem extended to include three events corresponding to the selection of ELTs from the three manufacturers (A, B, C):

$$P(A|D) = \frac{P(A)\cdot P(D|A)}{[P(A)\cdot P(D|A)]+[P(B)\cdot P(D|B)]+[P(C)\cdot P(D|C)]}$$

$$= \frac{0.80\cdot 0.04}{[0.80\cdot 0.04]+[0.15\cdot 0.06]+[0.05\cdot 0.09]}$$

$$= 0.703 \text{ (rounded)}$$

Basic Probability

- *Product rule*: probability $P(A \wedge B)$ of a conjunction of two events A and B

$$P(A \wedge B) = P(A|B)P(B) = P(B|A)P(A)$$

- *Sum rule*: probability of a disjunction of two events A and B

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

- *Bayes theorem*: the posterior probability $P(h|D)$ of h given D

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- *Theorem of total probability*: if events $A_1, \ldots, A_n$ are mutually exclusive with $\sum_{i=1}^{n} P(A_i) = 1$, then

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

BAYES THEOREM AND CONCEPT LEARNING

- What is the relationship between Bayes theorem and the problem of concept learning?
  - Since Bayes theorem provides a principled way to calculate the posterior probability of each hypothesis given the training data, we can use it as the basis for a straightforward learning algorithm that calculates the probability for each possible hypothesis, then outputs the most probable.

Brute-Force Bayes Concept Learning

- In particular, assume the learner considers some finite hypothesis space **H** defined over the instance space **X**, in which the task is to learn some target concept ***c*** : X → **{0,1}.**
- As usual, we assume that the learner is given some sequence of training examples ***<< x 1,d1 >***. . . ***<xm, dm>>*** where ***xi*** is some instance from X and where ***di*** is the target value of ***xi*** (i.e., ***di*** = ***c(xi)).***
- To simplify the discussion in this section, we assume the sequence of instances ***(x1*** . . . ***xm)*** is held fixed, so that the training data D can be written simply as the sequence of target values D = ***(d1*** . . . ***dm)***
- We can design a straightforward concept learning algorithm to output the maximum a posteriori hypothesis, based on Bayes theorem, as follows:

BRUTE-FORCE MAP LEARNING algorithm

1. For each hypothesis h in H, calculate the posterior probability

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

2. Output the hypothesis h_{MAP} with the highest posterior probability

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}} P(h|D)$$

Lets consider following assumptions

1. The training data ***D*** is noise free (i.e., ***di*** = ***c(xi)).***
2. The target concept ***c*** is contained in the hypothesis space H
3. We have no a priori reason to believe that any hypothesis is more probable than any other.

what values should we specify for ***P(h)?***

$$P(h) = \frac{1}{|H|} \quad \text{for all } h \text{ in } H$$

What choice shall we make for ***P(D/h)?***

P(D/h) is the probability of observing the target values D = ***<d1 . . .dm>*** for the fixed set of instances ***<X1 . . . xm>,*** given a world in which hypothesis ***h*** holds

$$P(D|h) = \begin{cases} 1 \text{ if } d_i = h(x_i) \text{ for all } d_i \text{ in } D \\ \\ 0 \text{ otherwise} \end{cases} \qquad (6.4)$$

In other words, the probability of data D given hypothesis h is 1 if D is consistent with h, and 0 otherwise.

$$P(h|D) = \frac{0 \cdot P(h)}{P(D)} = 0 \text{ if } h \text{ is inconsistent with } D$$

The posterior probability of a hypothesis inconsistent with D is zero.

Now consider the case where h is consistent with D.

$$P(D) = \sum_{h_i \in H} P(D|h_i) P(h_i)$$

$$P(h|D) = \frac{1 \cdot \frac{1}{|H|}}{P(D)}$$

$$= \sum 1 \cdot \frac{1}{|H|} + \sum 0 \cdot \frac{1}{|H|}$$

To summarize, Bayes theorem implies that the posterior probability $P(h|D)$ under our assumed $P(h)$ and $P(D|h)$ is

$$P(h|D) = \begin{cases} \frac{1}{|VS_{H,D}|} & \text{if } h \text{ is consistent with } D \\ 0 & \text{otherwise} \end{cases} \tag{6.5}$$

where $|VS_{H,D}|$ is the number of hypotheses from H consistent with D.

MAP Hypotheses and Consistent Learners

- We will say that a learning algorithm is a ***consistent learner*** **provided** it outputs a hypothesis that commits zero errors over the training examples.
- we can conclude that *every consistent learner outputs a MAP hypothesis, if **we** assume a uniform prior probability distribution over H (i.e., P(hi) = P(hj) for all **i,** j), and if we assume deterministic, noise free training data (i.e., P(D/h) = 1 if D and h are consistent, and 0 otherwise).*
- for example, the concept learning algorithm FIND-S:
  - FIND-S searches the hypothesis space H from specific to general hypotheses, outputting a maximally specific consistent hypothesis (i.e., a maximally specific member of the version space).
  - Because FIND-S outputs a consistent hypothesis, we know that it will output a MAP hypothesis under the probability distributions *P(h)* and *P(D|h)* defined above.

MAXIMUM LIKELIHOOD AND LEAST-SQUARED ERROR HYPOTHESES

- consider the problem of learning a ***continuous-valued target function***-a problem faced by many learning approaches such as neural network learning, linear regression, and polynomial curve fitting.
- A straightforward Bayesian analysis will show that ***under certain assumptions any learning algorithm that minimizes the squared error between the output hypothesis predictions and the training data will output a maximum likelihood hypothesis.***

Learner L considers an instance space X and a hypothesis space H consisting of some class of real-valued functions defined over X (i.e., each h in H is a function of the form $h : X \rightarrow \Re$, where $\Re$ represents the set of real numbers).

The problem faced by L is to learn an unknown target function $f : X \rightarrow \Re$ drawn from H.

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}}\, p(D|h)$$
$$= \underset{h \in H}{\operatorname{argmax}} \prod_{i=1}^{m} p(d_i|h)$$
$$= \underset{h \in H}{\operatorname{argmax}} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2}$$

Maximize natural log of this instead...

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}} \sum_{i=1}^{m} \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2$$
$$= \underset{h \in H}{\operatorname{argmax}} \sum_{i=1}^{m} -\frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2$$
$$= \underset{h \in H}{\operatorname{argmax}} \sum_{i=1}^{m} -\left(d_i - h(x_i)\right)^2$$
$$= \underset{h \in H}{\operatorname{argmin}} \sum_{i=1}^{m} \left(d_i - h(x_i)\right)^2$$

MAXIMUM LIKELIHOOD HYPOTHESES FOR PREDICTING PROBABILITIES

- we derive an analogous criterion for a second setting that is common in neural network learning: ***learning to predict probabilities***.

Consider predicting survival probability from patient data

Training examples $\langle x_i, d_i \rangle$, where d_i is 1 or 0

Want to train neural network to output a *probability* given x_i (not a 0 or 1)

In this case can show

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}} \sum_{i=1}^{m} d_i \ln h(x_i) + (1 - d_i) \ln(1 - h(x_i))$$

Weight update rule for a sigmoid unit:

$$w_{jk} \leftarrow w_{jk} + \Delta w_{jk}$$

where

$$\Delta w_{jk} = \eta \sum_{i=1}^{m} h(x_i)(1 - h(x_i))(d_i - h(x_i))\ x_{ijk}$$

MINIMUM DESCRIPTION LENGTH PRINCIPLE

Occam's razor: prefer the shortest hypothesis

The Minimum Description Length principle is motivated by interpreting the definition of h_{MAP} in the light of basic concepts from information theory. Consider again the now familiar definition of h_{MAP}.

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}} P(D|h)P(h)$$

which can be equivalently expressed in terms of maximizing the $\log_2$

$$h_{MAP} = \underset{h \in H}{\operatorname{argmax}} \log_2 P(D|h) + \log_2 P(h)$$

or alternatively, minimizing the negative of this quantity

$$h_{MAP} = \underset{h \in H}{\operatorname{argmin}} -\log_2 P(D|h) - \log_2 P(h)$$

- $-\log_2 P(h)$ is the description length of h under the optimal encoding for the hypothesis space H. In other words, this is the size of the description of hypothesis h using this optimal representation. In our notation, $L_{C_H}(h) = -\log_2 P(h)$, where C_H is the optimal code for hypothesis space H.
- $-\log_2 P(D|h)$ is the description length of the training data D given hypothesis h, under its optimal encoding. In our notation, $L_{C_{D|h}}(D|h) = -\log_2 P(D|h)$, where $C_{D|h}$ is the optimal code for describing data D assuming that both the sender and receiver know the hypothesis h.

Notice the expected length for transmitting one message is therefore $\sum_i -p_i \log_2 p_i$, the formula for the *entropy*

$$h_{MAP} = \underset{h}{\operatorname{argmin}}\, L_{C_H}(h) + L_{C_{D|h}}(D|h)$$

where C_H and $C_{D|h}$ are the optimal encodings for H and for D given h, respectively.

Minimum Description Length principle: Choose h_{MDL} where

$$h_{MDL} = \underset{h \in H}{\operatorname{argmin}} L_{C_1}(h) + L_{C_2}(D|h)$$

The above analysis shows that if we choose C_1 to be the optimal encoding of hypotheses C_H, and if we choose C_2 to be the optimal encoding $C_{D|h}$, then $h_{MDL} = h_{MAP}$.

$\rightarrow$ prefer the hypothesis that minimizes

$$length(h) + length(misclassifications)$$

Thank You!!!