innovate achieve lead

BITS Pilani

Machine Learning (IS ZC464) Session 4:

Bayes' Theorem and its applications in Machine Learning, MAP hypothesis, Information Theory and its application in Minimum Description Length (MDL) principle

Bayes' theorem

- Bayes' theorem provides a way to calculate the probability of a hypothesis based on its prior probability, the probability of observing various data given the hypothesis, and the observed data itself.

Bayesian learning
Example 1: observation of sounds

Training with Observed data: $\{d_1, d_2, d_3\}$ =training data (say D)

d_1: cat sounds with ‘ae’

d_2: pot sounds with ‘aw’

d_3: mat sounds with ‘ae’

Sounds ‘ae’ and ‘aw’ are the observed targets that we know.

Prior probabilities
P(sound = ‘ae’) = 0.5
P(sound = ‘aw’) = 0.5

features such as ‘a’ and ‘o’ are obtained through preprocessing of the given words – by parsing

Conditional probabilities are represented as
P(‘ae’ |feature = ‘a’)=2/3
P(‘aw’| feature = ‘o’) = 1/3

OR
P(‘ae’ |d1,d2,d3)=2/3
P(‘aw’|d1,d2,d3)= 1/3

OR
P(‘ae’ |D)=2/3
P(‘aw’|D)= 1/3

Bayesian learning would enable answers to queries such as:

Unknown words used for testing: sat and not

Preprocessing gives

Feature for word ‘sat’ = ‘a’

Feature for woed ‘not’ = ‘o’

What is the likelihood that word sat sounds with ‘ae’?

P(sat | ‘ae’) = ?

What is the likelihood that word not sounds with ‘ae’?

P(not | ‘ae’) = ?

What is the likelihood that word sat sounds with ‘aw’?

P(sat | ‘aw’) = ?

What is the likelihood that word not sounds with ‘aw’?

P(not | ‘aw’) = ?

Hypothesis

- In learning algorithms, the term hypothesis is used in contexts such as
  - Concept learning or classification: class label or category
  - Function approximation: a curve, a line or a polynomial
  - Decision making: a decision tree
- Plural of hypothesis: Hypotheses (multiple labels, multiple curves, multiple decision trees)
- Best Hypothesis (Always preferred) : Most appropriate class, best fit curve, smallest decision tree

Observations of Sounds example : continued

- There are two hypotheses in the given example
  - Hypothesis (say h_1) : 'ae'
  - Hypothesis (say h_2) : 'aw'
- **Learning:** requires us to find the best hypothesis from the space of two hypothesis h_1 and h_2 for a new observation

Likelihood or probability?

What is the likelihood that word 'sat' sounds with 'ae'?

P(sat | 'ae')

Which can equivalently be written as

P(feature = 'a' |'ae') = 2/3 (given)

Reverse:

What is the likelihood that 'ae' sound will represent a word of type sat?

P('ae'|feature = 'a') =?

Computation of P('ae'|feature = 'a') =?

- Let us represent the above probabilistic query as conditional probability using following events

  A: sound is 'ae'

  B: feature is 'a'

To compute P(A|B) (read as Probability of A given B) when P(B|A), P(B) and P(A) are available

Bayes' theorem provides a way to compute such probabilities

Terminology for Bayes' theorem

- **<u>Prior Probability:</u>** The probability P(h) denotes the initial probability that hypothesis 'h' holds before we have observed the training data. [Example: P('aw'), P('ae'), P(feature = 'a'), P(feature = 'o') etc. based on some background knowledge]

Terminology for Bayes' theorem

- **<u>Posterior Probability:</u>** The probability P(h |D) denotes the probability that the hypothesis 'h' holds given the observed training data D [First recall example of sequence of tossing of coin and the probability that changes as we keep observing the D. Then in the current example, P(feature = 'a'|'ae'), visualize uncertainty if 'talk' and 'none' are also used for training and sound different for feature = 'a' and feature = 'o' respectively]

Bayesian Learning

- Bayesian learning is a probabilistic approach to inference.
- Optimal decisions can be made by reasoning about these probabilities together with the observed data.
- Each observed training observation can incrementally decrease or increase the estimated probability that a hypothesis is correct.

Bayesian Learning

- 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.

Example

observation	word	feature	Target hypothesis (h)
d_1	put	u	oo
d_2	pat	a	ae
d_3	none	o	a~
d_4	mat	a	ae
d_5	cut	u	a~
d_6	not	o	aw
d_7	nut	u	a~
d_8	talk	a	aaw
d_9	pot	o	aw
d_{10}	sat	a	ae

4 hypotheses

Observations (D)	word	feature	Target hypothesis (h)
d_1	put	u	oo
d_2	pat	a	ae
d_3	none	o	a~
d_4	mat	a	ae
d_5	cut	u	a~
d_6	not	o	aw
d_7	nut	u	a~
d_8	talk	a	aw
d_9	pot	o	aw
d_{10}	sat	a	ae

Prior Probabilities
P(oo) = 0.1
P(ae) = 0.5
P(a~) = 0.1
P(aw)=0.3

conditionalProbabilities
P(oo | D) = 0.1
P(ae | D) = 0.3
P(a~ | D) = 0.3
P(aw | D)=0.3

Three features

Observations (D)	word	feature	Target hypothesis (h)
d_1	put	u	oo
d_2	pat	a	ae
d_3	none	o	a~
d_4	mat	a	ae
d_5	cut	u	a~
d_6	not	o	aw
d_7	nut	u	a~
d_8	talk	a	aw
d_9	pot	o	aw
d_{10}	sat	a	ae

Prior Probabilities of features

$P(u) = 0.2$

$P(a) = 0.5$

$P(o)= 0.3$

Conditional Probabilities

$P(u \mid oo) = 0.1$

$P(u \mid a\sim)= 0.2$

$P(a \mid ae) = 0.3$

$P(a \mid aw) =0.1$

$P(o \mid a\sim) = 0.1$

$P(o \mid aw)=0.2$

Bayesian Learning

- Training : Through the computation of the probabilities as in previous two slides
- Testing : of unknown words
  - Example testing:

  Which sound does the word ‘cat’ make?

  Preprocess(cat) to get feature ‘a’ and compute P(h|a), where P(h|D) is known, where D is the set of 10 observations used to train the system and ‘h’ is the hypothesis.

  Compute probabilities P(ae | a), P(oo | a), P(a~|a) and P(aw |a) to obtain the likelihood of sound of cat.

Posterior Probabilities

- P(ae |a) = P(a | ae) * P(ae)/P(a)

  = 0.3* 0.5/0.5 = **0.3**

Maximum

- P(oo |a) = P(a | oo) * P(oo)/P(a)

  = 0.1* 0.1/0.5 = **0.02**

- P(a~ |a) = P(a | a~) * P(a~)/P(a)

  = 0* 0.1/0.5 = **0**

- P(aw |a) = P(a | aw) * P(aw)/P(a)

  = 0.1* 0.3/0.5 = **0.06**

Conditional Probabilities
P(u | oo) = 0.1
P(u | a~)= 0.2
P(a |ae) = 0.3
P(a |aw) =0.1
P(o| a~) = 0.1
P(o | aw)=0.2

Prior Probabilities
P(oo) = 0.1
P(ae) = 0.5
P(a~) = 0.1
P(aw)=0.3

Prior Probabilities of features
P(u) = 0.2
P(a) = 0.5
P(o)= 0.3

Maximum a Posteriori (MAP) hypothesis

- Consider a set of hypotheses H and the observed data used for training D
- Define

$$h_{MAP} = \underset{h \in H}{Arg\max}\, P(h \mid D)$$

- The maximally probable hypothesis is called a maximum a posteriori (MAP) hypothesis.

MAP hypothesis

$$h_{MAP} = \underset{h \in H}{Arg\max}\, P(h \mid D)$$

$$h_{MAP} = \underset{h \in H}{Arg\max}\, \frac{P(D \mid h)P(h)}{P(D)}$$

using Bayes' theorem

$$h_{MAP} = \underset{h \in H}{Arg\max}\, P(D \mid h)P(h)$$

Dropping P(D) as id constant

Equally Probable hypothesis a priori

- If $P(h_i) = P(h_j)\ \forall\ h_i$ and h_j in H

  then in finding the MAP hypothesis, we can ignore the term P(h) in the following equation

$$h_{MAP} = \underset{h \in H}{Arg \max} P(D \mid h) P(h)$$

And get

$$h_{MAP} = \underset{h \in H}{Arg \max} P(D \mid h)$$

Maximum Likelihood hypothesis

- P(D|h) is called the likelihood of the data given h and any hypothesis that maximizes P(D|h) is called a Maximum Likelihood (ML) hypothesis.

$$h_{ML} = \underset{h \in H}{Arg \max} \, P(D \mid h)$$

Home Work

- Read and solve Example given in section 6.2.1 (Mitchell's book)
- Question on Bayes' theorem

  A doctor knows that the disease meningitis causes the patient to have a stiff neck, say 50% of the time. The doctor also knows some unconditional facts: the prior probability that the patient has meningitis is 1/50,000, and the prior probability that any patient has a stiff neck is 1/20. What is the probability that a patient with stiff neck has meningitis? (Verify your answer with 0.0002)

Minimum Description Length Principle

- This is the information theoretic approach to compute the MAP hypothesis.
- The MAP is computed as shortest length hypothesis in the domain of encoding data.
- A problem consisting of transmitting random messages needs encoding of messages.
- Messages are arriving at random (uncertainty about the messages exists)
- Each message 'i' is considered to be arriving with probability p_i

Minimum Description Length Principle

- We need to find the encoding scheme using minimum number of bits.
- The fixed length coding scheme does not work well as the less probable messages get the encoding using same number of bits.
- Example messages

  a_1, a_2, a_3, a_4 : 4 symbols

  Code them as 00, 01, 10, 11 using 2 bit (costly) representation

  Transmit the code sequence 1101100110001110.

  Client at the other end can decode using the same scheme of encoding as $a_4a_2a_3a_2a_3a_1a_4a_3$

Information Theory

- Information theory studies the quantification, storage, and communication of information.
- It was originally proposed by Claude E. Shannon in 1948 to find fundamental limits on signal processing and communication operations such as data compression
- A key measure in information theory is "entropy".
- Entropy quantifies the amount of uncertainty involved in the value of a random variable or the outcome of a random process.

Reference : https://en.wikipedia.org/wiki/Information_theory

Entropy

- Based on the probability of each source symbol to be communicated, the Shannon entropy *H*, in units of bits (per symbol), is given by

$$Entropy = -\sum_i p_i \log_2(p_i)$$

- where p_i is the probability of occurrence of the i^{th} possible value of the source symbol.

Encoding

- Less number of bits to represent frequent symbols
- Use more bits to represent less frequent symbols

symbol	**Probability (p_i)**	**Code (unoptimized)**	**Code (Optimized)**
a1	0.4	00	1
a2	0.25	01	010
a3	0.3	10	00
a4	0.05	11	011

Computation of entropy

- Entropy of the given data

$$Entropy = -\sum_i p_i \log_2(p_i)$$

$$Entropy = -p_1 \log_2(p_1) - p_2 \log_2(p_2) - p_3 \log_2(p_3) - p_4 \log_2(p_4)$$

= - 0.4*$\log_2$(0.4) - 0.25*$\log_2$(0.25)- 0.3*$\log_2$(0.3) - 0.05*$\log_2$(0.05)

= 0.52877 + 0.5 + 0.521089 + 0.216096

=**1.76** bits (information content)