COMPSCI762: Foundations of Machine Learning

Introduction to Bayesian Learning

Jörg Simon Wicker and Katerina Taškova

The University of Auckland

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

SCIENCE
SCHOOL OF COMPUTER SCIENCE

Bayesian Learning

Maximum Likelihood and Least-Squared Error
Minimum Description Length

Partly based on Mitchel's book, lecture slides from Stanford's NLP lecture and The University of Utah

Maximum Likelihood and Least-Squared Error

Maximum Likelihood and Least-Squared Error

- Problem: learning continuous-valued target functions (e.g. neural networks, linear regression, etc.)
- Bayesian analysis will show that under certain assumptions **any learning algorithm that minimizes the squared error between the hypothesis predictions and the training data, will output a maximum likelihood hypothesis.**

Maximum Likelihood and Least-Squared Error

- Problem setting:
  - Given a data set D containing m **training examples** of the form $< x_i, d_i >$
  - Let's say there exists **an unknown function** $f : X \to \mathbb{R}$ that describes how exactly the features from the input space X map to the target value defined over the set of real numbers $\mathbb{R}$
  - Given a hypothesis space $H : (\forall h \in H)[h : X \to \mathbb{R}]$, our goals is to find **the best hypothesis** $h*$ **that approximates** f.
  - Now assume the target value of each example is corrupted by **random noise** drawn independently according to a Normal probability distribution with zero mean $d_i = f(x_i) + e_i, e_i \sim Normal(0, \sigma^2)$

Maximum Likelihood and Least-Squared Error

$$h_{ML} = \arg\max_{h \in H} p(D|h)$$

- The training examples are assumed to be mutually independent given h

$$h_{ML} = \arg\max_{h \in H} \prod_{i=1}^{m} p(d_i|h)$$

- Given the noise e_i obeys a Normal distribution with mean $\mu = 0$ and unknown variance σ^2, each d_i must also obey a Normal distribution around the true target value $f(x_i)$. Hence, $\mu = f(x_i) = h(x_i)$

$$h_{ML} = \arg\max_{h \in H} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(d_i - h(x_i))^2}{2\sigma^2}}$$

Maximum Likelihood and Least-Squared Error

- How to find the best h^* from the previous equation?
  - We often compute log-likelihood instead of likelihood to make computation easier!
  - log() is a monotonically non-decreasing function, taking log of the likelihood does not affect the choice of the most probable hypothesis

$$h_{ML} = \arg\max_{h \in H} \sum_{i=1}^{m} log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(d_i - h(x_i))^2}{2\sigma^2}$$

- The first term in this expression is a constant independent of h and can therefore be discarded.

$$h_{ML} = \arg\max_{h \in H} \sum_{i=1}^{m} - \frac{(d_i - h(x_i))^2}{2\sigma^2}$$

- Maximizing this negative term is equivalent to minimizing the corresponding positive term.

$$h_{ML} = \arg\min_{h \in H} \sum_{i=1}^{m} \frac{(d_i - h(x_i))^2}{2\sigma^2}$$

Maximum Likelihood and Least-Squared Error

- Finally, all constants independent of h can be discarded.

$$h_{ML} = \arg\min_{h \in H} \sum_{i=1}^{m} (d_i - h(x_i))^2$$

⇒ the h_{ML} is one that minimizes the sum of the squared errors

- Why is it reasonable to choose the Normal distribution to characterize noise?
  - Good approximation of many types of noise in physical systems
  - Central Limit Theorem shows that the sum of a sufficiently large number of independent, identically distributed random variables itself obeys a Normal distribution
- Only noise in the target value is considered, not in the attributes describing the instances themselves

Minimum Description Length

Minimum Description Length Principle

- Occam's razor: choose the shortest explanation for the observed data
- Here, we consider a Bayesian perspective on this issue and a closely related principle
- Minimum Description Length (MDL) Principle
  - Motivated by interpreting the definition of h_{MAP} in the light of information theory concepts

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} P(D|h)P(h) \\ &= \arg\max_{h \in H} log_2 P(D|h) + log_2 P(h) \\ &= \arg\min_{h \in H} -log_2 P(D|h) - log_2 P(h) \end{aligned}$$

  - This equation can be interpreted as a statement that short hypotheses are preferred, assuming a particular representation scheme for encoding hypotheses and data

Minimum Description Length Principle

- Introduction to a basic result of information theory
  - Consider the problem of designing a code C to transmit messages drawn at random
  - Probability of encountering message i is p_i
  - Interested in the most compact code C
  - Shannon and Weaver (1949) showed that the optimal code assigns $-log_2 p_i$ bits to encode message i
  - $L_C(i) \approx$ description length of message i with respect to C

Minimum Description Length Principle

$$h_{MAP} = \arg\min_{h \in H} -log_2 P(D|h) - log_2 P(h)$$

- Interpret the equation using information theory
  - $L_{C_H}(h) = -log_2 P(h)$, where C_H **is the optimal code for hypothesis space** H
  - $L_{C_{D|h}}(D|h) = -log_2 P(D|h)$, where $C_{D|h}$ **is the optimal code for describing data** D assuming that both the sender and receiver know hypothesis h

  - $\Rightarrow$ Minimum description length principle

$$h_{MAP} = \arg\min_{h \in H} L_{C_H}(h) + L_{C_{D|h}}(D|h)$$

Minimum Description Length Principle

- To apply this principle in practice, **specific encodings or representations** appropriate for the given learning task must be chosen
- Application to decision tree learning
  - C_H might be some obvious encoding, in which the description length grows with the **number of nodes** and with the **number of edges**
  - Choice of $C_{D|h}$?
    - Assume both the sender and receiver know the sequences of m instances $< x_1, \ldots, x_m >$
    - What message do we need to transmit under this assumption?

1. If h correctly predicts the classification, no transmission is necessary $(L_{C_{D|h}}(D|h) = 0)$
2. In case of missclassification, for each missclassified instance a message has to be sent with the **id of the instance (at most $log_2 m$ bits)** as well as its **correct class label (at most $log_2 k$ bits, where k is the number of possible classes)**

Minimum Description Length Principle

- MDL principle provides a way for trading off hypothesis complexity for the number of errors committed by the hypothesis

  C_H : number-of-nodes + number-of-edges ⇒ **model complexity**
  $C_{D|h}$: $(log_2 m + log_2 k)\cdot$ number-of-missclassifications ⇒ **model errors**

  The shorter C_H is for a hypothesis, the more likely we make mistakes, and hence $C_{D|h}$ might be larger

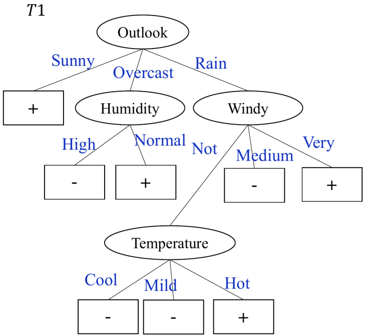
- One way of dealing with the issue of overfitting

MDL Example – Decision Tree Pruning

ID	Outlook	Temp.	Humidity	Windy	Class
I1	Overcast	Hot	High	Not	-
I2	Sunny	Mild	Normal	Very	+
...	...	...	... ...	...	
I32	Rain	Hot	High	Medium	-

Let's say we encode the tree with each row denoting a split. We can use 2 bits to encode the attribute and 1 bit to record a leaf node, e.g.

- Outlook: +, Humidity, Windy
- Humidity: -



100% accuracy

$$L_{C_{D|h}}(D|h) = 0$$

$$L_{C_H}(h) = \#\text{leaf} + 2\#\text{internal} = 8 + 6 = 14$$

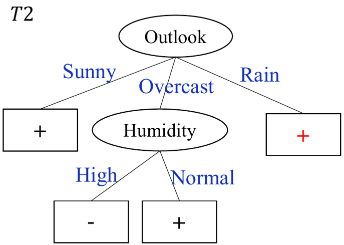
$$L_{C_{D|h}}(D|h) + L_{C_H}(h) = 0 + 14 = 14 \text{ bits}$$

MDL Example – Decision Tree Pruning

ID	Outlook	Temperature	Humidity	Windy	Class
I1	Overcast	Hot	High	Not	-
I2	Sunny	Mild	Normal	Very	+
…	…	…	… …	…	
I32	Rain	Hot	High	Medium	-

Let’s say we encode the tree with each row denoting a split. We can use 2 bits to encode the attribute and 1 bit to record a leaf node, e.g.

- Outlook: +, Humidity, Windy
- Humidity: -



Assume T2 missclassified only I32

$$L_{C_{D|h}}(D|h) = log_2 32 + log_2 2 = 5 + 1 = 6$$
$$L_{C_H}(h) = \#\text{leaf} + 2\#\text{internal} = 4 + 2 = 6$$
$$L_{C_{D|h}}(D|h) + L_{C_H}(h) = 6 + 6 = 12 \text{ bits}$$

Summary

- Bayesian learning relies on Bayes' Theorem
- Bayesian methods can be used to select the most likely hypothesis (MAP/ML) given the data
- Bayesian Learning has multiple roles
  - Provide practical and effective learning algorithms like Naive Bayes
  - Provide a framework
    - For evaluating other learners
    - For analyzing learning
- Bayes optimal classifier combines the predictions of all alternative hypothesis weighted by their posterior probabilities
- Bayesian networks provide a natural representation for conditional independence
- Naive Bayes classifier is a simple and fast method for classification that assumes attribute values are conditionally independent given the target value.

Literature

- Chapter 6 of Mitchell's *Machine Learning* (also look at Section 2 of `www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf`)
- Chapter 8 of Bishop's *Pattern Recognition and Machine Learning*

Thank you for your attention!

`https://ml.auckland.ac.nz`