

Machine Learning I: Bayesian learning

Ch. 18.1-2, 20.1-2

Wed Nov 5, 2008

COMP131 Lecture 15

1

What we know

- How to represent uncertain knowledge with Bayes Nets
- How to do probabilistic inference
- Where do the structure and conditional probabilities of Bayes nets come from?
 - expert knowledge
 - learned from data/experience
- We've seen how to learn a policy for acting in a world with unknown dynamics with reinforcement learning
- Now we'll see how to learn other kinds of models when we have more or less data of different kinds

Wed Nov 5, 2008

COMP131 Lecture 15

2

Today

- Machine learning I (of II)
 - the problem
 - types of learning
 - hypothesis spaces
 - noise and overfitting
 - Bayesian learning

Wed Nov 5, 2008

COMP131 Lecture 15

3

The machine learning problem

- Propose a general rule based on examples:
induction

Piece of bread 1 was nourishing when I ate it
 Piece of bread 2 was nourishing when I ate it
 ...
 Piece of bread 100 was nourishing when I ate it

Therefore, all pieces of bread will be nourishing if I eat them



David Hume: 1711-1776

Wed Nov 5, 2008

COMP131 Lecture 15

4

Machine learning problems

- Classification
 - is this an indoor or outdoor visual scene?
 - is this person a friend or an enemy?
 - should the applicant with this credit history be given a mortgage or not?
- Regression
 - describe a function (mapping) from input features to some output space
 - what's a good control law (mapping from sensor data to position or velocity motor commands) for a mobile robot?
- Causal explanation
 - what is the causal relationship between events in the world?
- Prediction
 - what will happen next? what will happen if I do this?

Wed Nov 5, 2008

COMP131 Lecture 15

5

Types of learning

- What kind of information is available for the learning agent?
- Examples of <inputs, desired output>: **supervised learning**
- Examples of <inputs>: **unsupervised learning**
- Examples of <inputs, actions, rewards>: **reinforcement learning**

Wed Nov 5, 2008

COMP131 Lecture 15

6

Toy problem: classification

Supervised learning:
examples are labeled
with desired class output

Wed Nov 5, 2008 COMP131 Lecture 15 7

Hypothesis space

- Supervised learning problem: given data of the form $\langle x, f(x) \rangle$, find the function f , called a hypothesis
- Discrete f : classification
- Continuous f : regression
- Can use f to make predictions about future inputs

- Need to choose H – the hypothesis class

Wed Nov 5, 2008 COMP131 Lecture 15 8

Toy problem: regression

- $H = \{\text{polynomials up to degree } 10\}$
- Perfect fit: sinusoidal

Wed Nov 5, 2008 COMP131 Lecture 15 9

Occam's razor

- “Entities should not be multiplied unnecessarily” (William of Ockham c.1285-1349)
- Noise-complexity (overfitting) tradeoff:
maximize $E(h, D) + \alpha C(h)$
- No free lunch theorem: any hypothesis that fits your data is as good as any other, unless you already know something about the kinds of problems you're likely to encounter in the future

Wed Nov 5, 2008 COMP131 Lecture 15 10

Candy problem

- Surprise candy comes in two flavors: cherry (yum) and lime (yuck), wrapped in identical opaque wrappers. Candy is sold in very large bags indistinguishable from the outside, of which there are five types:
 - h1: 100% cherry
 - h2: 75% cherry, 25% lime
 - h3: 50% cherry, 50% lime
 - h4: 25% cherry, 75% lime
 - h5: 100% lime
- The manufacturer advertises that the distribution over bag types is $\langle .1, .2, .4, .2, .1 \rangle$
- Agent can open and inspect N pieces of candy from a bag

Wed Nov 5, 2008 COMP131 Lecture 15 11

Bayesian learning

- Calculate the probability of each hypothesis given data
 $P(h_i | d) = \alpha P(d | h_i) P(h_i)$
- To predict the probability distribution over some unknown quantity X
 $P(X | d) = \sum_i P(X | h_i) P(h_i | d)$
- If the observations d are i.i.d., then
 $P(d | h_i) = \prod_j P(d_j | h_i)$
- Suppose the first 10 candies we taste are all lime
 $P(d | h_3) = 0.5^{10} \approx 0.001$

Wed Nov 5, 2008 COMP131 Lecture 15 12

As number of samples increases

Maximum A Posteriori (MAP) prediction:
predict from just the most likely hypothesis

Dangerous! But much simpler than summing over a large H

Wed Nov 5, 2008 COMP131 Lecture 15 13

A natural bias against complexity

- Bayesian learning asks for a prior distribution over hypotheses
- Can encode bias against complex hypotheses by setting their prior probabilities low
- Bayesian Occam's razor (MAP: simplest theory that is consistent with the data)
 - this result is based on information theory; the MAP hypothesis is also the minimum description length hypothesis

Wed Nov 5, 2008 COMP131 Lecture 15 14

Uniform priors

- If nothing is known about prior hypothesis distributions, assume a uniform prior (all hypotheses equally likely)
- Simplified MAP gives Maximum Likelihood (ML) hypothesis:
 h_{ML} maximizes $P(h_i | \mathbf{d}) = \alpha P(\mathbf{d} | h_i)$

Wed Nov 5, 2008 COMP131 Lecture 15 15

Example

- A new manufacturer of candy doesn't advertise prior distributions. Here's a bag of candy. It has some proportion θ of cherry candy in it (so $1-\theta$ lime).
- Hypothesis space $H = [0, 1]$, each h_θ hypothesizes a particular proportion of cherry candy
- We open N candies from the bag, c of which are cherry and $l=N-c$ are lime. What's the likelihood of this data set given some h_θ ?
- $P(\mathbf{d} | h_\theta) = \prod_j P(d_j | h_\theta) = \theta^c (1-\theta)^l$
- Taking the log of the likelihood and the derivative wrt θ (see derivation AIMA p.716-718), setting it to zero gives the parameter θ which maximizes likelihood

Wed Nov 5, 2008 COMP131 Lecture 15 16

Maximum likelihood learning

- Write down an expression for the likelihood of data given hypothesis as a function of hypothesis parameter
- Take the derivative of the log likelihood with respect to the parameter
- Find the parameter values such that the derivative is zero

In many cases, 3 is tricky and will require iterative solutions or numerical optimization methods (stochastic search).

Another problem with ML: when data set small, some events will have occurred 0 times, and the ML hypothesis assigns 0 probability to those events...

Wed Nov 5, 2008 COMP131 Lecture 15 17

Naïve Bayes learning

- Problems: the "class" variable is root, and the "attribute" variables are leaves
- Assume attributes are conditionally independent of each other given the class (naïve, but powerful)
- Assuming Boolean variables, the parameters are:
 $\theta = P(C=true)$, $\theta_{i1} = P(X_i=true|C=true)$, $\theta_{i2} = P(X_i=true|C=false)$
- Train parameters with max. log likelihood of data
- The classify with the probability for each class for a new example: ...

Wed Nov 5, 2008 COMP131 Lecture 15 18

Taking advantage of Bayes net

- Need continuous-valued random variables (nodes in Bayes net)
- Replace sums with integrals
- Learning parameters is the same as inference

Wed Nov 5, 2008

COMP131 Lecture 15

19

Learning structure?

- A Bayes Net is specified by a DAG with parameters (cond. prob. tables)
- Can search for a good Bayes net with a scoring function that trades off
 - maximization of likelihood of data given net
 - complexity of net

Wed Nov 5, 2008

COMP131 Lecture 15

20

Learning with hidden variables?

- Suppose you are given data about patients' smoking, diet, exercise, chest pain, fatigue and shortness of breath

Wed Nov 5, 2008

COMP131 Lecture 15

21

Summary

- Machine learning is the problem of generalizing a dataset into a function (discrete: classification, continuous: regression)
- Bayesian learning requires prior distributions over the hypothesis space H , predicts based on all hypotheses
- The MAP hypothesis gives approximately Bayesian predictions (better with more samples)
- The ML hypothesis works with uniform priors, using log likelihood of data to estimate parameters
- Bayes net parameters can be learned with inference on continuous-valued nodes

Wed Nov 5, 2008

COMP131 Lecture 15

22

Next time

- Neural networks

Wed Nov 5, 2008

COMP131 Lecture 15

23