Lecture 3: Bayesian Learning

- Bayes Theorem
- Most likely hypotheses: MAP and ML
- Using Bayes theorem to understand regression
- Bayes optimal classifier

Questions

- How can we represent and understand prior knowledge and bias?
- What is a reasonable bias?
- What sorts of error measures should we use in classification and regression?
- When are different error measures appropriate?
- What is an optimal classifier?

Two roles for Bayesian theory

1. Provides useful conceptual framework
   - "Gold standard" for evaluating other learning algorithms
   - Provides rationale for other aspects of learning taken for granted, such as Occam's razor, use of mean-squared error etc.
2. Provides practical learning algorithms:
   - Naive Bayes learning
   - Bayesian belief network learning (not covered in this class)

   which combine prior knowledge (prior probabilities) with observed data

Random variables and probability

- A **random variable** X describes an outcome that cannot be determined in advance (e.g. the roll of a die)
- The **sample space** S of a random variable X is the set of all possible values of the variable

  E.g. For a die, $S = \{1, 2, 3, 4, 5, 6\}$
- An **event** is a subset of S. E.g. $e = \{1\}$ corresponds to a die roll of 1
- Usually, random variables are still governed by some law of nature, described as a **probability function** p defined on S. $p(x)$ defines the chance that variable X takes value $x \in S$.

  E.g. for a die roll with a fair die, $p(1) = p(2) = \ldots = p(6) = \frac{1}{6}$

  Note: We still cannot determine the value of X, just the chance of encountering a given value

Discrete random variables

If X is a discrete variable, then a probability space $p(x)$ has the following properties:

$$0 \leq p(x) \leq 1, \forall x \in S \text{ and } \sum_{x \in S} p(x) = 1$$

Axioms of probability

For any events (propositions) A, B:

1. $0 \leq P(A) \leq 1$
2. $P(True) = 1$
3. $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$, or equivalently, $P(A \vee B) = P(A) + P(B)$ if A and B are mutually exclusive

The axioms of probability limit the class of functions that can be considered probability functions.

Using functions that disobey these laws as probabilities can force suboptimal decisions (de Finetti, 1931).

Conditional probability

The basic statements in the Bayesian framework talk about **conditional probabilities**. $P(A|B)$ is the belief in event A given that event B is known with absolute certainty:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \text{ if } P(B) \neq 0$$

Note that we can use either the set intersection or the logical "and" notation above.

The **product rule** gives an alternative formulation:

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

Bayes rule

Bayes rule is another alternative formulation of the product rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

The **complete probability formula** states that:

$$P(A) = P(A|B)P(B) + P(A|\neg B)P(\neg B)$$

or more generally,

$$P(A) = \sum_i P(A|b_i)P(b_i),$$

where b_i form a set of exhaustive and mutually exclusive events.

Example: Using Bayes theorem

A patient takes a lab test and the result comes back positive. 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. Furthermore, $.008$ of the entire population have this cancer. Does patient have cancer or not?

$$P(cancer) = \qquad P(\neg cancer) =$$

$$P(+|cancer) = \qquad P(-|cancer) =$$

$$P(+|\neg cancer) = \qquad P(-|\neg cancer) =$$

$$P(cancer|+) =$$

Bayes theorem in learning

Let h be a hypothesis and D be the set of training data. Using Bayes theorem, we have:

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

where:

- $P(h)$ = 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

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

What is the most probable hypothesis given the training data?

Maximum a posteriori (MAP) hypothesis h_{MAP}:

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

Choosing hypotheses (continued)

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

If we assume $P(h_i) = P(h_j)$ (all hypotheses are equally likely a priori) then can further simplify, and choose the **maximum likelihood (ML) hypothesis**:

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

Brute force MAP hypothesis learner

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

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

Relation to concept learning

Consider our usual concept learning task: instance space X, hypothesis space H. Assume a fixed set of examples $D = \{\langle x_1, c_1 \rangle, \ldots, \langle x_m, c_m \rangle\}$.

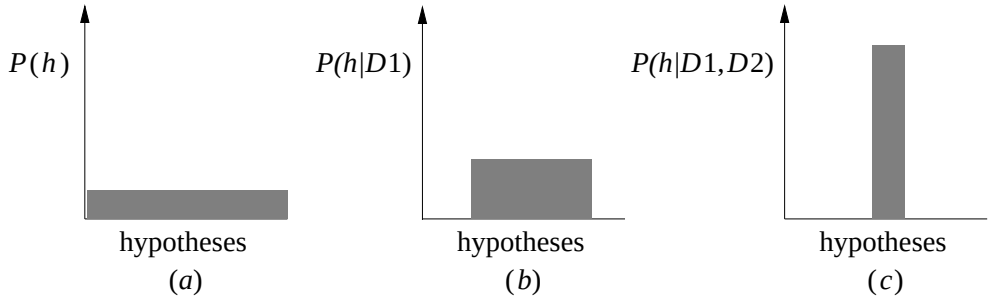
Assuming perfect data, we have:

$$P(D|h) = \begin{cases} 1 & \text{if } h \text{ consistent with} D \\ 0 & \text{otherwise} \end{cases}$$

Choose $P(h)$ to be *uniform* distribution: $P(h) = \frac{1}{|H|}$ for all h in H

What are the MAP hypotheses in this case?

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



Using Bayes theorem to understand regression

Consider the task of learning a real-valued target function f.

The training examples are $\langle x_i, d_i \rangle$, where d_i is a noisy target value:

$$d_i = f(x_i) + e_i,$$

where e_i is random variable (noise) drawn independently for each x_i according to some Gaussian distribution with zero mean

What is the maximum likelihood hypothesis?

ML hypothesis for regression with Gaussian noise

$$
\begin{aligned}
h_{ML} &= \arg\max_{h \in H} p(D|h) \\
&= \arg\max_{h \in H} \prod_{i=1}^{m} p(d_i|h) \quad \text{(because the data points are independent)} \\
&= \arg\max_{h \in H} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2} \quad \text{(because the noise is Gaussian)}
\end{aligned}
$$

ML hypothesis for regression with Gaussian noise (2)

Standard trick: if you have to deal with products, take the log of it, and you get a nice sum:

$$
\begin{aligned}
h_{ML} &= \arg\max_{h \in H} \sum_{i=1}^{m} \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2 \\
&= \arg\max_{h \in H} \sum_{i=1}^{m} -\frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2 \\
&= \arg\max_{h \in H} \sum_{i=1}^{m} -\left(d_i - h(x_i)\right)^2 \\
&= \arg\min_{h \in H} \sum_{i=1}^{m} \left(d_i - h(x_i)\right)^2
\end{aligned}
$$

ML hypothesis for regression with Gaussian noise (3)

Then the maximum likelihood hypothesis h_{ML} is the one that minimizes our old friend, **the sum of squared errors**:

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

This makes explicit the hypothesis behind minimizing the sum-squared error: independent data examples, with Gaussian target noise

Most probable classification of new instances

MAP and ML give the most probable *hypothesis* given the data D

Given new instance x, what is its most probable *classification*?

$h_{MAP}(x)$ (also called the **Naive Bayes classification** is **NOT** the most probable classification!

Example:

Consider three possible hypotheses:

$$P(h_1|D) = .4,\ P(h_2|D) = .3,\ P(h_3|D) = .3$$

Given a new instance x,

$$h_1(x) = 1,\ h_2(x) = 0,\ h_3(x) = 0$$

What is the most probable classification of x?

Bayes Optimal Classifier

Bayes optimal classification:

$$\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

Note that this classifier uses hypotheses in H, but may not be in H itself!

Example

In our example:

$$P(h_1|D) = .4, \quad P(-|h_1) = 0, \quad P(+|h_1) = 1$$
$$P(h_2|D) = .3, \quad P(-|h_2) = 1, \quad P(+|h_2) = 0$$
$$P(h_3|D) = .3, \quad P(-|h_3) = 1, \quad P(+|h_3) = 0$$

Therefore

$$\sum_{h_i \in H} P(+|h_i)P(h_i|D) = .4$$
$$\sum_{h_i \in H} P(-|h_i)P(h_i|D) = .6$$

and the most probably classification is 0.

Gibbs Classifier

Bayes optimal classifier provides best result, but can be expensive if many hypotheses.

Gibbs algorithm:

1. Choose one hypothesis at random, according to $P(h|D)$
2. Use this to classify new instance

Surprising fact: Assume target concepts are drawn at random from H according to correct priors on H. Then:

$$E[error_{Gibbs}] \leq 2E[error_{BayesOptimal}]$$

Example

Suppose correct, uniform prior distribution over all hypotheses consistent with the training data

Then, according to Gibbs:

- Pick any hypothesis from the version space, with uniform probability
- Its expected error no worse than twice Bayes optimal!

Summary

- Bayes theorem!!!
- General view of learning algorithms:
  - MAP: maximize $P(h|D)\alpha P(D|h)P(h)$
  - ML: maximize $P(D|H)$
  - Bayes optimal: produces the most likely classification
- Regression with Gaussian noise: minimizing the squared error