Lecture 14: Learning - Introduction and Parameter Estimation

- Why learning?
- Learning problems in Bayes nets
- Learning with complete data
  - Maximum likelihood estimation
  - Bayesian estimation

Why learning?

- Isn't this obvious??? 8-)
- Knowledge acquisition bottleneck
  - Experts are scare and expensive, often inconsistent
  - In some cases, experts do not exist!
  - But data is cheap and plentiful!
- The main goal of learning is to build probabilistic "expert systems" based on data

Learning in Bayesian networks

Given data, in the form of **instances**, e.g.:

Earthquake	Burglary	Alarm	Call	Radio
No	No	No	No	No
No	Yes	Yes	Yes	No
...				

Create a complete Bayes net (graph structure + CPDs)

Two dimensions

1. Known vs. unknown structure of the network
   I.e. do we know the arcs or do we have to find them too?
2. Complete vs. incomplete data
   In real data sets, values for certain variables may be missing, and we need to "fill in"

Today we focus on the easy case: known network structure, complete data

This problem is called **parameter estimation**

Parameter estimation with complete data in directed models

- Given:
  - a network structure G
  - a choice of representation for $p(X_i|\text{Parents}(X_i))$ (e.g. a table, in the case of discrete variables)
- Goal: learn the CPDs in each node, such that the network is "closest" to the probability distribution that generated the data
- For now, assume that we have binary variables

Statistics 101: Bernoulli trials

- We have a coin that can land in two positions (heads or tails)
- Let $p(H) = \theta$ be the unknown probability of the coin landing head
- Given a sequence of tosses $x_1, x_2, \ldots x_m$ we want to **estimate** $p(H)$.

More generally: Statistical parameter fitting

- Given instances $x_1, \ldots x_m$ that are **i.i.d.**:
  - The set of possible values for each variable in each instance is known
  - Each instance is sampled from the same distribution
  - Each instance is independent of the other instances
- Find a set of parameters θ such that the data can be summarized by a probability $p(x_i|\theta)$
- θ depends on the family of probability distributions we consider (e.g. multinomial, Gaussian etc.)

How good is a parameter set?

- It depends on how likely it is to generate the observed data
- Let D be the data set (all the instances)
- The **likelihood** of parameter set θ given data set D is defined as:

$$L(\theta|D) = p(D|\theta)$$

- If the instances are i.i.d., we have:

$$L(\theta|D) = p(D|\theta) = \prod_{j=1}^{m} p(x_j|\theta)$$

- E.g. In the coin tossing problem, the likelihood of a parameter θ given the sequence $D = H, T, H, T, T$ is:

$$L(\theta|D) = \theta(1-\theta)\theta(1-\theta)(1-\theta) = \theta^{N_H}(1-\theta)^{N_T}$$

Sufficient statistics

- To compute the likelihood in the coin tossing example, we only need to know N_H and N_T (number of heads and tails)
- We say that N_H and N_T are **sufficient statistics** for the binomial distribution
- In general, a sufficient statistic of the data is a function of the data that summarizes enough information to compute the likelihood
- Formally, $s(D)$ is a sufficient statistic if, for any two data sets D and D',

$$s(D) = s(D') \Rightarrow L(\theta|D) = L(\theta|D')$$

Maximum likelihood estimation (MLE)

- **Choose parameters that maximize the likelihood function**
- We want to maximize:

$$L(\theta|D) = \prod_{j=1}^{m} p(x_j|\theta)$$

  This is a product, and products are hard to maximize!
- Standard trick: maximize $\log L(\theta|D)$ instead!

$$\log L(\theta|D) = \sum_{i=1}^{m} \log p(x_j|\theta)$$

- To maximize, we take the derivatives of this function with respect to θ and set them to 0

MLE applied to the binomial data

- The likelihood is:

$$L(\theta|D) = \theta^{N_H}(1-\theta)^{N_T}$$

- The log likelihood is:

$$\log L(\theta|D) = N_H \log\theta + N_T \log(1-\theta)$$

- Take the derivative of the log likelihood and set it to 0:

$$\frac{\partial}{\partial\theta}\log L(\theta|D) = \frac{N_H}{\theta} + \frac{N_T}{1-\theta}(-1) = 0$$

- Solving this gives

$$\theta = \frac{N_H}{N_H + N_T}$$

- This is intuitively appealing!

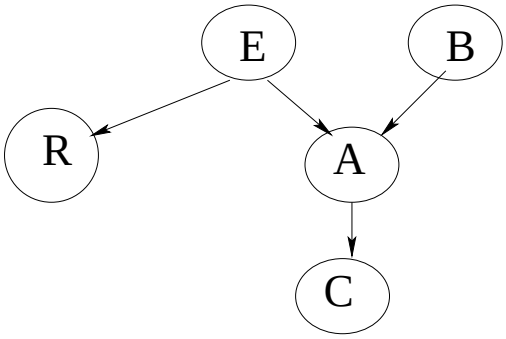
Observations

- Depending on our choice of probability distribution, when we take the gradient of the likelihood we may not be able to find θ analytically
- An alternative is to do **gradient descent** instead:
  1. Start with some guess $\hat{\theta}$
  2. Update $\hat{\theta}$:

$$\hat{\theta} \leftarrow \hat{\theta} + \alpha\frac{\partial}{\partial\theta}\log L(\theta|D)$$

     where $\alpha \in (0,1)$ is a **learning rate**
  3. Go back to 2 (for some number of iterations, or until θ stops changing significantly
- Sometimes we can also determine a **confidence interval** around the value of θ

Parameter estimation in a Bayes net: Example



- Instances are of the form $\langle r_j, e_j, b_j, a_j, c_j \rangle, j = 1, \ldots m$
- What are the parameters we are trying to estimate?

Example (continued)

$$
\begin{aligned}
L(\theta|D) &= \prod_{j=1}^{m} p(r_j, e_j, b_j, c_j, a_j|\theta) \text{ (from i.i.d)} \\
&= \prod_{j=1}^{m} p(e_j)p(r_j|e_j)p(b_j)p(a_j|e_j, b_j)p(c_j|e_j) \text{ (factorization)} \\
&= (\prod_{j=1}^{m} p(e_j))(\prod_{j=1}^{m} p(r_j|e_j))(\prod_{j=1}^{m} p(b_j))(\prod_{j=1}^{m} p(a_j|e_j, b_j))(\prod_{j=1}^{m} p(c_j|e_j)) \\
&= \prod_{i=1}^{n} L(\theta_i|D)
\end{aligned}
$$

where θ_i are the parameters associated with node i.

Parameter estimation for general Bayes nets

Generalizing, for any Bayes net with variables $X_1, \ldots X_n$, we have:

$$
\begin{aligned}
L(\theta|D) &= \prod_{j=1}^{m} p(X_1(j), \ldots X_n(j)|\theta) \text{ (from i.i.d)} \\
&= \prod_{j=1}^{m} \prod_{i=1}^{n} p(X_i(j)|Parents(X_i(j)), \theta) \text{ (factorization)} \\
&= \prod_{i=1}^{n} \prod_{j=1}^{m} p(X_i(j)|Parents(X_i(j)) \\
&= \prod_{i=1}^{n} L(\theta_i|D)
\end{aligned}
$$

The likelihood function **decomposes** according to the structure of the network, which creates **independent estimation problems**.

Consistency of MLE

- For any estimator, we would like the parameters to converge to the "best possible" values as the number of examples grows
  We need to define "best possible" for probability distributions
- Let P and Q be two probability distributions over X. The **Kullback-Leibler divergence** between p and q is defined as:

$$KL(p, q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

- MLE is a **consistent estimator**, in the sense that (under a set of standard assumptions), w.p.1, we have:

$$\lim_{|D| \to \infty} \theta = \theta^*,$$

where θ^* is the "best" set of parameters:

$\theta^* = \arg\min_\theta KL(p^*(X), p(X|\theta))$ (p^* is the true distribution)

Summary: MLE (frequentist) approach

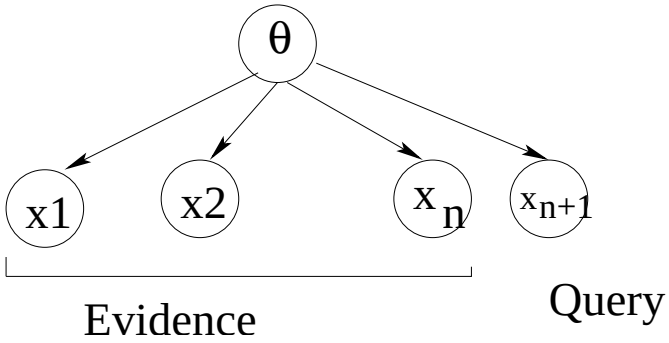
- Assume there is an unknown, fixed set of parameters θ
- Estimate θ (possibly with some confidence measure)
- Predict future events based on the estimated parameter value

Is this all we need?

- Suppose you observed 10 coin tosses, and 7 came up heads. Would you bet on heads in the next toss?
- What if instead of a coin you toss a ball painted in two different colors?

Bayesian approach

- The unknown parameters are represented as **random variables**
- We represent the uncertainty in the sampling process itself using a Bayes net



- Now prediction is just inference in this network

Prediction as inference

$$\begin{aligned} p(x_{n+1}|x_1, \ldots x_n) &= \int p(x_{n+1}|\theta, x_1, \ldots x_n) p(\theta|x_1, \ldots x_n) d\theta \\ &= \int p(x_{n+1}|\theta) p(\theta|x_1, \ldots x_n) d\theta, \end{aligned}$$

where

$$p(\theta|x_1, \ldots x_n) = \frac{p(x_1, \ldots x_n|\theta) p(\theta)}{p(x_1 \ldots x_n)}$$

Note that $p(x_1 \ldots x_n)$ is just a normalizing factor and $p(x_1, \ldots x_n|\theta) = L(\theta|D)$.

Example: Binomial data

- Suppose we observe 1 toss, $x_1 = H$. What would the MLE be?
- In the Bayesian approach,

$$p(\theta|x_1, \ldots x_n) \propto p(x_1, \ldots x_n|\theta) p(\theta)$$

- Assume we have a uniform prior for $\theta \in [0, 1]$, so $p(\theta) = 1$ (remember that θ is a continuous variable!)
- Then we have:

$$\begin{aligned} p(x_2 = H|x_1 = H) &\propto \int_0^1 p(x_1 = H|\theta) p(\theta) p(x_2 = H|\theta) d\theta \\ &= \int_0^1 \theta \cdot 1 \cdot \theta = \frac{1}{3} \end{aligned}$$

Example (continued)

- Likewise, we have:

$$\begin{aligned} p(x_2 = T|x_1 = H) &\propto \int_0^1 p(x_1 = H|\theta)p(\theta)p(x_2 = T|\theta)d\theta \\ &= \int_0^1 \theta \cdot 1 \cdot (1-\theta) = \frac{1}{6} \end{aligned}$$

- By normalizing we get:

$$\begin{aligned} p(x_2 = H|x_1 = H) &= \frac{\frac{1}{3}}{\frac{1}{3} + \frac{1}{6}} = \frac{2}{3} \\ p(x_2 = T|x_1 = H) &= \frac{1}{3} \end{aligned}$$

- It is as if we had our original data, plus two more tosses! (one heads, one tails)
- Suppose now that we get another toss, $x_2 = T$. What is $p(X_3|x_1 = H, x_2 = T)$?

Prior knowledge

- The prior incorporates prior knowledge or beliefs about the parameters
- As data is gathered, these beliefs do not play a significant role anymore
- More specifically, if the prior is well-behaved (does not assign 0 probability to feasible parameter values), MLE and Bayesian approach both give consistent estimators, so they converge in the limit to the same answer
- But the MLE and Bayesian predictions typically differ after fixed amounts of data. But in the short run, the prior can impact the **speed** of learning!
- More on why next time...