CMSC828G Principles of Data Mining Lecture #7

- Announcements:
  - Eiman's office hours changed to T 12:30-1:30PM, Th 9:30-10:30AM
  - Matlab review today 2PM, AVW 3228
- Today's Reading:
  - HMS, chapter 4
- Today's Lecture:
  - Student presentations
  - Parameter Estimation
- Upcoming Due Dates:
  - H1 due 2/21

Recall

- Independence: P(X,Y) = P(X)P(Y)
- Conditional Independence: P(X,Y|Z) = P(X|Z)P(Y|Z)
- Example 1: I(X,Y|∅) and not I(X,Y|Z)
- Example 2: I(X,Y|Z) and not I(X,Y|∅)
- conclusion: independence does not imply conditional independence or vice-versa!

Importance of Independence

- Independence/Conditional Independence are important properties to identify in a distribution
- They provide one criterion on which a joint distribution can be ***factored*** into a product of simpler marginal and/or conditional distributions

Sequential Data

- sequence of values, $x_1, \ldots, x_n$
- a common assumption made is that next value in the sequence is independent of all of the past values given the current value:

$$P(X_j \mid X_{j-1} \ldots X_1) = P(X_j \mid X_{j-1})$$

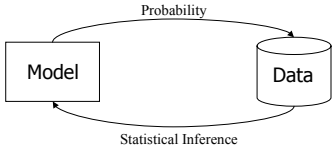
- **first-order Markov assumption**
- Allows factorization of joint into simpler products:

$$p(x_1, \ldots, x_n) = p(x_1) \prod_{j=2}^{n} p(X_j \mid X_{j-1})$$

Samples

- In data mining, sometimes we work with entire population of interest
- Other times we work with a sample from the population
- Even if the entire data set is available, we may work with a sample:
  - entirely legitimate if we are interested in learning a model
  - may be less appropriate when we are looking for patterns of anomalous behavior

Statistical Inference



- Statistical inference: inferring properties of an unknown distribution from data generated by that distribution.
- most common: approximating the unknown distribution by choosing a distribution from a restricted family of distributions.
- **Estimate** parameters of the model from data
- (if we had the entire population, we would **calculate** parameters)

Statistical Inference, cont.

- assumes that the sample has been drawn from the population at random
- The model specifies the distribution for the population; the probability that a particular value for the variable will appear in the sample
- If we have a model M for the data, we can compute the probability that a random sampling process will lead to the data D = {x(1),…,x(x)}:

$$p(D \mid M)$$

- if we assume the probability of each data point is independent, or 'drawn at random' then

$$p(D \mid \theta, M) = \prod_{i=1}^{n} p(x(i) \mid \theta, M)$$

Statistical Inference, cont.

$$p(D \mid \theta, M) = \prod_{i=1}^{n} p(x(i) \mid \theta, M)$$

- based on this value, we can decide how realistic the assumed model is.
- if the assumed model is unlikely to have generated the data, we might reject the model; this is the principle behind hypothesis testing
- we can also estimate population values for the parameters…

Desirable properties of estimators

- Let $\hat{\theta}$ be an estimate for a population parameter θ
- The value we compute for $\hat{\theta}$ depends on the data
- For different samples, we will have different estimates
- in other words, $\hat{\theta}$ is a random variable
- it will have a distribution, mean, $E(\hat{\theta})$ and $Var(\hat{\theta})$
- The **bias** of an estimator is defined:

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

- An estimator is **unbaised** if

$$E(\hat{\theta}) = \theta$$

- i.e., no systematic departure from the true parameter on average

Desirable estimator properties, cont.

- The variance of an estimator is another measure of quality.
- The variance of an estimator is:

$$Var(\hat{\theta}) = E\left[(\hat{\theta} - E(\hat{\theta}))^2\right]$$

- Measures how sensitive the estimator is to individual data sets
- Choose between estimators that have the same bias by choosing one with minimum variance
- unbiased estimators with minimum variance are called **best unbiased estimators**

Example

- Suppose we ignore data D and simply say that $\hat{\theta}$ is 1.0
- $Var(\hat{\theta})$ is 0
- however, in most cases this estimator will have a bias that will be nonzero and large

Bias-Variance decomposition of MSE

- The mean squared error (MSE) of $\hat{\theta}$ is

$$E\left[(\hat{\theta} - \theta)^2\right] = E\left[\left(\hat{\theta} - E\left[\hat{\theta}\right] + E\left[\hat{\theta}\right] - \theta\right)^2\right]$$
$$= \underbrace{(E\left[\hat{\theta}\right] - \theta)^2}_{\text{Bias}^2} + \underbrace{E\left[(\hat{\theta} - E\left[\hat{\theta}\right])^2\right]}_{\text{Variance}}$$

- MSE is useful criteria, since it measures systematic bias and random variance between estimate and true value
- Unfortunately, bias and variance often work in different directions; reducing an estimator's bias tends to increase the variance and vice-versa.

BIAS VARIANCE TRADEOFF

More desirable properties

- Let $\hat{\theta}_{n_1}, \hat{\theta}_{n_2}, \ldots, \hat{\theta}_{n_m}$ be a sequence of estimators based on increasing sample sizes n_1, n_2, …, n_m
- The sequence is consistent if

$$\lim_{n\to\infty} \text{pr}(\left|\hat{\theta}_n - \theta\right| > \varepsilon) = 0$$

Parameter estimation

- Maximum Likelihood Estimation
- Bayesian Estimation

Likelihood Function

- Let D = {x(1),…,x(n)}
- independently sampled, from the same distribution p(x|θ)
  'independent and identically distributed', iid
- The **likelihood function**, L(θ| x(1),…,x(n)) captures the probability of the data as a function of θ

$$\begin{aligned} L(\theta, D) &= L(\theta \mid x(1),\ldots, x(n)) \\ &= p(x(1),\ldots, x(n) \mid \theta) \\ &= \prod_{i=1}^{n} p(x(i) \mid \theta) \end{aligned}$$

Maximum Likelihood Estimation (MLE)

- Most widely used method of parameter estimation
- The likelihood function, L(θ| x(1),…,x(n))

$$\begin{aligned} L(\theta, D) &= L(\theta \mid x(1),\ldots, x(n)) \\ &= p(x(1),\ldots, x(n) \mid \theta) \\ &= \prod_{i=1}^{n} p(x(i) \mid \theta) \end{aligned}$$

- choose value that maximizes the likelihood function

$$\hat{\theta}_{\text{MLE}}$$

Example

- Database of customer purchases, want to estimate probability that a randomly chosen customer buys milk
- Suppose we have random sample 1000 customers that either do or do not buy milk, D ={x(1), …,x(n)}
- Assume simple Binomial model where θ is the probability that milk is purchased

$$L(\theta \mid x(1),\ldots, x(1000)) = \prod_{i=1}^{1000} \theta^{x(1)}(1-\theta)^{1-x(1)} = \theta^{r}(1-\theta)^{1000-r}$$

- Take logs:

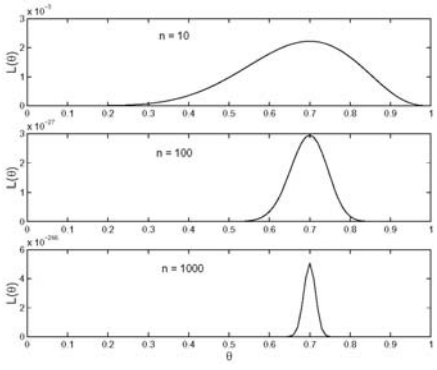
$$l(\theta) = r\log\theta + (1000-r)\log(1-\theta)$$

- Differentiate and set to zero:

$$\frac{r}{\theta} - \left(\frac{1000-r}{1-\theta}\right) = 0$$

$$\hat{\theta}_{\text{MLE}} = \frac{r}{1000}$$

Binomial Likelihood Function



Next Time

- Reading:
  - HMS, chapter 4 cont.
- Topic:
- Due:
  - homework #1

References

- *Principles of Data Mining*, Hand, Mannila, Smyth. MIT Press, 2001.
- http://www.cc.gatech.edu/classes/cs6751_97_winter/Topics/stat-meas/probHist.html
- http://mathforum.org/epigone/apstat-l/folkixblon