Markov Models

Assume the next request depends only on the last request.

Much more powerful than purely random probability.

- Ex: There is a distribution of letters in English, but it is different for beginning letters in a word

Transition matrices

- Ex: typical distribution for some system is about 4 small packets followed by a single large packet.
- Random chance: probability of small is always .8, large is always .2.
- Markov model: Small follows small .75, large follows small .25. In contrast, small always follows large.
Clustering

Goal: To separate a population into groups with similar characteristics.
- Specifically, the goal is to minimize the with-group variance while maximizing the between-group variance.

Uses:
- Select representatives to simplify further processing
- Recommender systems, retrieval systems

Steps:
- (Random) sampling
- Parameter selection
- Transformation of parameters (log scale?), scale (normalize data)
- Cull outliers
- Select distance metric (e.g., euclidean distance)
- Clustering technique. Ex: minimum spanning tree
- Assign full population to clusters
Variables Types

Variables can be classified into types, with increasing ability to perform mathematical calculations

Categorical variables
- Nominal values, no rank ordering possible
- ex: Sex (male or female)

Ordinal measurement
- Rank is important, but cannot compare the values any further
- ex: Position in race (first, second, third), we have no idea whether second is close or far from first

Interval measurement
- Equal distance between units, but no absolute zero value so cannot take ratios
- ex: Temperature in Ferinhiet, 60 degrees is 30 units more than 30 degrees, but its not twice as hot

Ratio Scales
- There is a fixed zero, so ratios make sense
- ex: One salary value can be twice another salary value
Statistics: Basic Concepts

Independent Events: Knowing that one event has occurred does not in any way change our estimate of the probability of the other event.

Random Variable: takes one of a specified set of values with specified probability.

Mean or Expected Value (for discrete variables):

\[ \mu = E(x) = \sum_{i=1}^{n} p_i x_i \]

Variance:

\[ \text{Var}(x) = \sigma^2 = E[(x - \mu)^2] = \sum_{i=1}^{n} p_i (x_i - \mu)^2 \]

Standard Deviation: \[ \sqrt{\text{Var}(x)} = \sigma \]

Coefficient of Variation: \[ \sigma / \mu \]
Basic Concepts (cont)

Covariance: Given random variables $x$ and $y$ with means $\mu_x$ and $\mu_y$:

$$\text{Cov}(x, y) = \sigma_{xy}^2 = E[(x - \mu_x)(y - \mu_y)] = E(xy) - E(x)E(y)$$

Correlation: Normalized value of covariance (always between -1 and 1)

$$\text{Correlation}(x, y) = \rho_{xy} = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y}$$

Alternate formulation (Pearson’s $r$):

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \mu_x}{\sigma_x} \right) \left( \frac{y_i - \mu_y}{\sigma_y} \right)$$
Mean, Median, Mode

Cumulative Distribution Function: The probability of a variable taking a value less than or equal to \( a \):

\[
F_x(a) = P(x \leq a)
\]

- Histogram

Quantile: Denoted \( x_\alpha \), the \( x \) value at which the CDF takes a value \( \alpha \) is called \( \alpha \)-quantile.

Mean or Expected Value: \( \mu = E(x) = \sum_{i=1}^{n} p_i x_i \)

Median: The 50-percentile (or 0.5-quantile)

Mode: The most likely value
Choosing Mean, Median, Mode

Often want to give a single number that “characterizes” a data set.

- Indices of Central Tendency
- Favorites are Mean, Median, Mode
- The term “average” has no meaning

Each has its appropriate use (and inappropriate uses)

You can’t take the median or mean of categorical data. Use Mode.

Does the total have value? (Ex: Total CPU time for five database queries.) Probably use mean.

Are the data skewed? Use Median
- GTA’s mistake
- Otherwise, can use mean or median
Indices of Dispersion

“Then there is the man who drowned crossing a stream with an average depth of six inches.”

Mean, median, mode attempt to provide a single “characteristic” value for the population.
- But a single value might not be meaningful
- Some populations have similar values through, others vary a lot
- People generally prefer systems whose response time is more consistent

There are various ways to measure variability

Range (max - min)
- poor, tends to be unbounded, unstable over a range of observations, susceptible to outliers

Variance ($s^2$), standard deviation ($s$)

\[
s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2
\]
- standard deviation is in units of the mean

10- and 90-precentiles

Quartiles (box plots)