Describing Data

- The canonical descriptive strategy is to describe the data in terms of their underlying distribution.
- As usual, we have a p-dimensional data matrix with variables $X_1, \ldots, X_p$.
- The joint distribution is $P(X_1, \ldots, X_p)$.
- The joint gives us complete information about the variables.
- Given the joint distribution, we can answer any question about the relationships among any subset of variables:
  - are $X_2$ and $X_5$ independent?
  - generating approximate answers to queries for large databases or selectivity estimation.
- Given a query (conditions that observations must satisfy), estimate the fraction of rows that satisfy this condition (the selectivity of the query).
- These estimates are needed during query optimization.
- If we have a good approximation for the joint distribution of data, we can use it to efficiently compute approximate selectivities.

Difficulty: Curse of Dimensionality

- Consider joint distribution for multivariate categorical data.
- $p$ variables, each taking $m$ values.
- The joint distribution requires specifying $O(mp)$ different probabilities.
- Exponential # of parameters is problem for:
  - estimation
  - representation and reasoning.

CoD: Estimation

- We can think of $mp$ cells, $(c_1, \ldots, c_{mp})$ each containing $n_i$ observations.
- The expected number of data points in cell $i$, given a random sample from $p(x)$ of size $n$ is $E_{p(x)}[n_i]=nmp$.
- Suppose $p(x)$ is uniform, i.e., $p(x) = 1/mp$.
- Then $E_{p(x)}[n_i] = \frac{n}{mp}$.
- If $n < 0.5mp$ then the expected number of points in any cell is closer to 0 than 1.
- If we use MLE estimator, $p_i=0$ for each empty cell.
- Fundamental problem:
  - If we have a data set of size $n$ over $p$ variables and we want to increase the number of variables from $p$ to $2p$, in order to keep the expected number of data points in each cell the same, we must increase the size of the data set by a factor of $mp$!!

CoD: Reasoning

- Even if we can reliably estimate a full joint distribution from the data, it is exponential in both space and time to manipulate it directly.
- For example if we want to determine the marginal distribution of any single variable $x_i$, we calculate it as follows:
  $$p(x_i) = \sum_{x_{i+1}, \ldots, x_p} p(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p)$$
- This requires summing over all the other variables and requires $O(m^n)$ summations.
- Working directly with the full joint distribution is feasible for only relatively low-dimensional problems.

Models

- Parametric
- Nonparametric
- Mixture distributions
- Semi-parametric
Parametric Models
- assume a particular, relatively simple, functional form
- e.g., uniform distribution, normal distribution, exponential, Poisson
- typically relatively small number of parameters
- often closed form solutions for parameter estimates that require a single pass through the data
- important to test the assumptions made by the model:
  - using simple visualizations
  - using statistical goodness-of-fit tests

Nonparametric Models
- take a local data-driven weighted average of around the point of interest
- simplest version: histogram
  - estimate for density is just (scaled) number of points in bin
  - problems:
    - not smooth
    - choosing the number of bins, bin locations and widths
  - ok for large data sets, small p
- regardless, generally useful to look at histograms with large number of bins, since can provide info on outliers, multimodality, skewness, tail behavior, etc.

Nonparametric Models cont.
- kernel density estimates
- density at any point x is proportional to a weighted sum of all points in the training data set
- weights are defined by an appropriately defined kernel function
- in 1-D:
  \[ f(x) = \frac{1}{n} \sum w_i \cdot K \left( \frac{x - x(i)}{h} \right) \]
  \[ K(t) = 1 - |t|, \quad t \leq 1; \quad K(t) = 0 \text{ otherwise} \]
  h determines smoothness; many possible choices for K
- we’ll talk about this more when we discuss SVMs....

Mixture Distributions
- Assume a probability model for each component
- Mixture Model:
  \[ f(x) = \sum_{k=1}^{K} w_k f_k(x; \theta_k) \]
  - distribution is linear combination of simpler distributions
  - where \( f_k \) are component distributions
  - components: gaussian, poisson, exponential
  - unlike simple parametric models, typically no closed form solution for maximizing score; one approach: use EM

Semi-parametric Models
- general class of functional forms in which the number of adaptive parameters can be increased in a systematic way to build ever more flexible models, but where the total number of parameters in the model can be varied independently from the size of the data set. Bishop
- e.g., mixture-models (where \( k \) varies), neural networks, graphical models.

Graphical Models
- In the next 3-4 lectures, we will be studying graphical models
  - e.g. Bayesian networks, Bayes nets, Belief nets, Markov networks, etc.
  - We will study:
    - representation
    - reasoning
    - learning
  - Materials based on upcoming book by Nir Friedman and Daphne Koller. Slides courtesy of Nir Friedman.
Probability Distributions

- Let \( X_1, \ldots, X_p \) be random variables
- Let \( P \) be a joint distribution over \( X_1, \ldots, X_p \)

If the variables are binary, then we need \( O(2^p) \) parameters to describe \( P \)

Can we do better?
- **Key idea:** use properties of independence

Independent Random Variables

- Two variables \( X \) and \( Y \) are independent if
  - \( P(X = x | Y = y) = P(X = x) \) for all values \( x, y \)
  - That is, learning the values of \( Y \) does not change prediction of \( X \)
- If \( X \) and \( Y \) are independent then
  - \( P(X,Y) = P(X|Y)P(Y) = P(X)P(Y) \)
- In general, if \( X_1, \ldots, X_p \) are independent, then
  - \( P(X_1, \ldots, X_p) = P(X_1) \ldots P(X_p) \)
  - Requires \( O(n) \) parameters

Conditional Independence

- Unfortunately, most of random variables of interest are not independent of each other
- A more suitable notion is that of conditional independence

Two variables \( X \) and \( Y \) are conditionally independent given \( Z \) if
  - \( P(X = x | Y = y, Z = z) = P(X = x | Z = z) \) for all values \( x, y, z \)
  - That is, learning the values of \( Y \) does not change prediction of \( X \) once we know the value of \( Z \)
  - notation: \( I (X, Y | Z) \)

Example: Naïve Bayesian Model

- A common model in early diagnosis:
  - Symptoms are conditionally independent given the disease (or fault)
- Thus, if
  - \( X_1, \ldots, X_p \) denote whether the symptoms exhibited by the patient (headache, high-fever, etc.) and
  - \( H \) denotes the hypothesis about the patient’s health
- then, \( P(X_1, \ldots, X_p, H) = P(H)P(X_1|H) \ldots P(X_p|H) \)
- This naïve Bayesian model allows compact representation
  - It does embody strong independence assumptions

Example: Family trees

Noisy stochastic process:

Example: Pedigree
- A node represents an individual’s genotype

Modeling assumptions:
  - Ancestors can effect descendants’ genotype only by passing genetic materials through intermediate generations

Example: Family trees

Markov Assumption

- We now make this independence assumption more precise for directed acyclic graphs (DAGs)
- Each random variable \( X_i \) is independent of its non-descendents, given its parents \( Pa(X) \)
- Formally,
  - \( I (X, NonDesc(X) | Pa(X)) \)
Markov Assumption Example

- In this example:
  - I(E, B)
  - I(B, E, R)
  - I(R, A, B, C | E)
  - I(A, R | B, E)
  - I(C, B, E, R | A)

Earthquake
Burglary
Alarm
Cell

I-Maps
- A DAG $G$ is an I-Map of a distribution $P$ if the all Markov assumptions implied by $G$ are satisfied by $P$
  (Assuming $G$ and $P$ both use the same set of random variables)

Examples:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$X$</td>
<td>$Y$</td>
<td>$P(x, y)$</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0.25</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.25</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.25</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Factorization
- Given that $G$ is an I-Map of $P$, can we simplify the representation of $P$?

- Example:

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$X$</td>
<td>$Y$</td>
<td>$P(x, y)$</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0.2</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.3</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.4</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Factorization Theorem

**Thm:** If $G$ is an I-Map of $P$, then

$$P(X_1, \ldots, X_p) = \prod_{i=1}^{p} P(X_i | Pa(X_i))$$

**Proof:**

- By chain rule:
  $$P(X_1, \ldots, X_p) = \prod_{i=1}^{p} P(X_i | X_1, \ldots, X_{i-1})$$
- wlog. $X_1, \ldots, X_p$ is an ordering consistent with $G$

From assumption: $Pa(X_i) \subseteq \{X_i, \ldots, X_p\}$

- Since $G$ is an I-Map, $I(X_i, NonDesc(X_i))$
- Hence:
  $$I(X_i, \{X_i, \ldots, X_{i-1}\} \rightarrow Pa(X_i))$$
  $$Pa(X_i) \subseteq NonDesc(X_i)$$

We conclude, $P(X_i | X_{i-1}, \ldots, X_1) = P(X_i | Pa(X_i))$

Consequences
- We can write $P$ in terms of “local” conditional probabilities

If $G$ is sparse,
- that is, $|Pa(X_i)| < k$,
  $$\Rightarrow$$
  each conditional probability can be specified compactly
  - e.g. for binary variables, these require $O(2^k)$ params.
  $$\Rightarrow$$ representation of $P$ is compact
  - linear in number of variables
Summary

- Probability distribution as descriptive model
- Difficulty: curse of dimensionality
- Categories of density estimation:
  - parametric
  - nonparametric
  - semi-parametric
- Graphical models tackle the curse of dimensionality by exploiting conditional independence

Next Time

- HW2 due

References

- Nir Friedman’s excellent lecture notes, [http://www.cs.huji.ac.il/~nmi/](http://www.cs.huji.ac.il/~nmi/)