Why build Models?

- To predict (identify) something
- Diagnosis
- Best therapy
- Prognosis
- Cost
- To understand something
  - Structure of model may correspond to structure of reality

Where do models come from?

- Pure induction from data
- Even so, need some “space” of models to explore
- Maximum A-posteriori Probability (MAP)
  \[ P(h_i|d) = \alpha P(d|h_i) P(h_i) \]
- Maximum Likelihood (ML)
  \[ P(h_i|d) = \alpha P(d|h_i) \]
  - Assumes uniform priors over all hypotheses in the space
- A-priori knowledge, expressed in
  - Structure of the space of models
  - \( P(h_i) \)
  - Adjustments to observed data

An Example

(Russell & Norvig)

- Surprise Candy Corp. makes two flavors of candy: cherry and lime
- Both flavors come in the same opaque wrapper
- Candy is sold in large bags, which have one of the following distributions of flavors, but are visually indistinguishable:
  - \( h_1 \): 100% cherry
  - \( h_2 \): 75% cherry, 25% lime
  - \( h_3 \): 50% cherry, 50% lime
  - \( h_4 \): 25% cherry, 75% lime
  - \( h_5 \): 100% lime
- Relative prevalence of these types of bags is (.1, .2, .4, .2, .1)
- As we eat our way through a bag of candy, predict the flavor of the next piece; actually a probability distribution.
Bayesian Learning

- Calculate the probability of each hypothesis given the data
  \[ P(h_i|d) = \alpha P(d|h_i)P(h_i) \]
- To predict the probability distribution over an unknown quantity, \( X \),
  \[ P(X|d) = \sum_{i} P(X|d, h_i)P(h_i|d) = \sum_{i} P(X|h_i)P(h_i|d) \]
- If the observations \( d \) are independent, then
  \[ P(d|h_i) = \prod_j P(d_j|h_i) \]
- E.g., suppose the first 10 candies we taste are all lime
  \[ P(d|h_3) = 0.5^{10} \approx 0.001 \]

Observations

- Bayesian approach asks for prior probabilities on hypotheses!
  - Natural way to encode bias against complex hypotheses: make their prior probability very low
- Choosing \( h_{MAP} \) to maximize \( P(h_i|d) = \alpha P(d|h_i)P(h_i) \)
  - is equivalent to minimizing \( -\log P(d|h_i) - \log P(h_i) \)
  - but as we know that entropy is a measure of information, these two terms are
    - # of bits needed to describe the data given hypothesis
    - # bits needed to specify the hypothesis
  - Thus, MAP learning chooses the hypothesis that maximizes compression of the data; Minimum Description Length principle
- Assuming uniform priors on hypotheses makes MAP yield \( h_{ML} \), the maximum likelihood hypothesis, which maximizes \( P(h_i|d) = \alpha P(d|h_i) \)

Learning Hypotheses and Predicting from Them

- (a) probabilities of \( h_i \) after \( k \) lime candies; (b) prob. of next lime
- MAP prediction: predict just from most probable hypothesis
  - After 3 limes, \( h_5 \) is most probable, hence we predict lime
  - Even though, by (b), it’s only 80% probable

Learning More Complex Hypotheses

- Input:
  - Set of cases, each of which includes
    - numerous features: categorical labels, ordinals, continuous
    - these correspond to the independent variables
- Output:
  - For each case, a result, prediction, classification, etc., corresponding to the dependent variable
  - In regression problems, a continuous output
  - a designated feature the model tries to predict
  - In classification problems, a discrete output
    - the category to which the case is assigned
- Task: learn function \( f(\text{input}) = \text{output} \)
  - that minimizes some measure of error
Linear Regression

- General form of the function
  \[ y = f(x_1, x_2, \ldots, x_n) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \]
- For each case:
  \[ \hat{y}_i = f(x_{1,i}, x_{2,i}, \ldots, x_{n,i}) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_n x_{n,i} \]
- Find \( \beta_i \) to minimize some function of \( y_i - \hat{y}_i \) over all \( y_i \)
  - e.g., mean squared error: \[ \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]

Logistic Regression

- Logistic function: \[ f(x) = \frac{1}{1 + e^{-x}} \]
- \[ y_i = f(x_i) \]
- \[ z_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_n x_{n,i} \]
- E.g., how risk factors contribute to probability of death
- \( \beta_i \) are the log odds ratios \( \log O(y_i|x_i) \)

More sophisticated models

- Nearest Neighbor Methods
- Classification Trees
- Artificial Neural Nets
- Support Vector Machines
- Bayes Networks (much on this, later)
- Rough Sets, Fuzzy Sets, etc. (see 6.873/HST951 or other ML classes)

How?

- Given: pile of training data, all cases labeled with gold standard outcome
- Learn “best” model
- Gather new test data, also all labeled with outcomes
- Test performance of model on new test data
- Simple, no?
Simplest Example

- Relationship between a diagnostic conclusion and a diagnostic test

<table>
<thead>
<tr>
<th></th>
<th>Test Positive</th>
<th>Test Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease Present</td>
<td>True Positive</td>
<td>False Negative</td>
</tr>
<tr>
<td>Disease Absent</td>
<td>False Positive</td>
<td>True Negative</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Test Positive</th>
<th>Test Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease Present</td>
<td>True Positive</td>
<td>False Negative</td>
</tr>
<tr>
<td>Disease Absent</td>
<td>False Positive</td>
<td>True Negative</td>
</tr>
</tbody>
</table>

\[TP+FP\] \[FN+TN\]

Definitions

<table>
<thead>
<tr>
<th></th>
<th>Test Positive</th>
<th>Test Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease Present</td>
<td>True Positive</td>
<td>False Negative</td>
</tr>
<tr>
<td>Disease Absent</td>
<td>False Positive</td>
<td>True Negative</td>
</tr>
</tbody>
</table>

\[TP+FN\] \[FP+TN\]

Sensitivity (true positive rate): \[\frac{TP}{TP+FN}\]

\[\text{False negative rate: } 1-\text{Sensitivity} = \frac{FN}{TP+FN}\]

Specificity (true negative rate): \[\frac{TN}{FP+TN}\]

\[\text{False positive rate: } 1-\text{Specificity} = \frac{FP}{FP+TN}\]

Positive Predictive Value (PPV): \[\frac{TP}{TP+FP}\]

Negative Predictive Value (NPV): \[\frac{TN}{FN+TN}\]
Test Thresholds Change Trade-off between Sensitivity and Specificity

Receiver Operator Characteristic (ROC) Curve

What makes a better test?

Need to explore many models

- Remember:
  - training set => model
  - model + test set => measure of performance
- But
  - How do we choose the best family of models?
  - How do we choose the important features?
  - Models may have structural parameters
    - Number of hidden units in ANN
    - Max number of parents in Bayes Net
  - Parameters (like the betas in LR), and meta-parameters
  - Not legitimate to “try all” and report the best !!!!!!!!!!!!!!!!
The Lady Tasting Tea

- R. A. Fisher & the lady
  - B. Muriel Bristol claimed she prefers tea added to milk rather than milk added to tea
  - Fisher was skeptical that she could distinguish
- Possible resolutions
  - Reason about the chemistry of tea and milk
    - Milk first: a little tea interacts with a lot of milk
    - Tea first: vice versa
  - Perform a “clinical trial”
    - Ask her to determine order for a series of cups
    - Calculate probability that her answers could have occurred by chance; if small, she “wins”
  - Fisher’s Exact Test
- Significance test
  - Reject the null hypothesis (that it happened by chance) if its probability is less than 0.10, 0.05, 0.01, 0.001, ...?

How to deal with multiple testing

- Suppose Ms. Bristol had tried this test 100 times, and passed once. Would you be convinced of her ability to distinguish?
- Bonferroni correction: for \( n \) trials, insist on a p-value that is \( 1/n \) of what you would demand for a single trial

Cross-validation

- Any number of times
  - Train on some subset of the training data
  - Test on the remainder, called the validation set
  - Choose best meta-parameters
  - Train, with those meta-parameters, on all training data
  - Test on Test data, once!

Aliferis lessons (part)

- Overfitting
  - bias, variance, noise
  - \( O \) = optimal possible model over all possible learners
  - \( L \) = best model learnable by this learner
  - \( A \) = actual model learned
  - \( \text{Bias} = O - L \) (limitation of learning method or target model)
  - \( \text{Variance} = L - A \) (error due to sampling of training cases)
  - Compare against learning from randomly permuted data
- Curse of dimensionality
  - Feature selection
  - Dimensionality reduction
Causality

- Suppes, 1950's
  - Statistical association
  - Temporal succession
  - No confounders (!)
  - hidden variables
- A node, $X$, is conditionally independent of all other nodes in the network given its Markov blanket: its parents, $U$, children $Y$, and children's parents, $Z$.

Cleaning the data—half the research time

- Missing values
  - Some values are not measured for some clinical situations
  - Failures in data capture process
- Episodically measured variables
- Unclear/undefined clinical states
- Imprecise timing of meds, ...
- Partially measured i/o
- Proxies: e.g., which ICU—what disease
- Derived variables: integrals, slopes, ranges, frequencies, etc.
- Transformed variables: square root, log, etc.
- Select subset of data with enough data!

Using MIMIC data to build predictive models

- Mortality
  - Comparison to SAPS II
  - Daily Acuity Scores
  - Real-time Acuity Scores

Descriptive look
If a patient leaves the hospital alive within this 30-day period, "mortality" excludes deaths that occur after long post-ICU stays. Only the first ICU stay of the first recorded hospital visit is counted as mortality. Work, references to "mortality" indicate death in the ICU or in the hospital for a significant period of time before dying.

The change in mortality rate as patients stay in the ICU for longer than seven days is plotted in Figure 3-6. By performing backward elimination one last time on this model selection (all development data) output for the day, it marks a patient as at-risk of mortality (negative correlation).

Table 3.14: Final Dataset: Partial Patient Exclusions

<table>
<thead>
<tr>
<th>Year</th>
<th>Start</th>
<th>End</th>
<th>Age</th>
<th>Gender</th>
<th>Type of admission</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>1/1</td>
<td>12/31</td>
<td>18</td>
<td>Female</td>
<td>Elective admission</td>
</tr>
</tbody>
</table>

Table 4.1: SAPS II Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Max Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>18</td>
</tr>
<tr>
<td>Heart rate</td>
<td>11</td>
</tr>
<tr>
<td>Systolic BP</td>
<td>13</td>
</tr>
<tr>
<td>Body temperature</td>
<td>3</td>
</tr>
<tr>
<td>PaO2/FiO2 (if ventilated or continuous)</td>
<td>11</td>
</tr>
<tr>
<td>Urinary output</td>
<td>11</td>
</tr>
<tr>
<td>Serum urea nitrogen level</td>
<td>10</td>
</tr>
<tr>
<td>WBC count</td>
<td>12</td>
</tr>
<tr>
<td>Serum potassium</td>
<td>3</td>
</tr>
<tr>
<td>Serum sodium level</td>
<td>5</td>
</tr>
<tr>
<td>Serum bicarbonate level</td>
<td>6</td>
</tr>
<tr>
<td>Bilirubin level</td>
<td>9</td>
</tr>
<tr>
<td>Glasgow Coma Scorea</td>
<td>26</td>
</tr>
<tr>
<td>Chronic diseases</td>
<td>17</td>
</tr>
<tr>
<td>Type of admission</td>
<td>8</td>
</tr>
</tbody>
</table>

*If the patient is sedated, the estimated GCS prior to sedation.

Many univariate analyses

SDAS Model Sensitivity

<table>
<thead>
<tr>
<th>Fold 1</th>
<th>Fold 2</th>
<th>Fold 3</th>
<th>Fold 4</th>
<th>Fold 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDAS</td>
<td>SDAS</td>
<td>SDAS</td>
<td>SDAS</td>
<td>SDAS</td>
</tr>
<tr>
<td>S.D.</td>
<td>S.D.</td>
<td>S.D.</td>
<td>S.D.</td>
<td>S.D.</td>
</tr>
</tbody>
</table>

SAPS II

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coef</th>
<th>S.E.</th>
<th>Wald</th>
<th>Z</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.531</td>
<td>0.045</td>
<td>3.38</td>
<td>0.0007</td>
<td></td>
</tr>
<tr>
<td>INR_mean_i</td>
<td>1.000</td>
<td>0.179</td>
<td>5.70</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>WBC_count</td>
<td>0.000</td>
<td>0.000</td>
<td>0.00</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>Sodium</td>
<td>0.011</td>
<td>0.006</td>
<td>1.97</td>
<td>0.0560</td>
<td></td>
</tr>
<tr>
<td>BUNtoCr</td>
<td>0.000</td>
<td>0.000</td>
<td>0.00</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>GCSrdv_mean</td>
<td>0.013</td>
<td>0.007</td>
<td>1.93</td>
<td>0.0540</td>
<td></td>
</tr>
<tr>
<td>GCS_max_sq</td>
<td>0.000</td>
<td>0.000</td>
<td>0.00</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>INR_mean_i</td>
<td>1.000</td>
<td>0.179</td>
<td>5.70</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>WBC_count</td>
<td>0.000</td>
<td>0.000</td>
<td>0.00</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>Sodium</td>
<td>0.011</td>
<td>0.006</td>
<td>1.97</td>
<td>0.0560</td>
<td></td>
</tr>
<tr>
<td>BUNtoCr</td>
<td>0.000</td>
<td>0.000</td>
<td>0.00</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>GCSrdv_mean</td>
<td>0.013</td>
<td>0.007</td>
<td>1.93</td>
<td>0.0540</td>
<td></td>
</tr>
<tr>
<td>GCS_max_sq</td>
<td>0.000</td>
<td>0.000</td>
<td>0.00</td>
<td>1.0000</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.15: Preprocessed Data

<table>
<thead>
<tr>
<th>Number of Patients</th>
<th>10,066</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Rows</td>
<td>150,496,922</td>
</tr>
<tr>
<td>Number of Features</td>
<td>438</td>
</tr>
</tbody>
</table>
### Evaluating the models

**Devel**

- ROC curve (development data):
  - AUC: 0.882
  - SDAS: All Days

**Validation**

- ROC curve:
  - AUC: 0.87
  - SDAS: Day 4
  - SDAS: Day 2

### Selected features for each day of ICU stay

- DAS model (day n)
  - Intercept
  - GCS_max
  - alloutput_max
  - BloodCal_max
  - platelets_max
  - platelets_slope
  - Resp_mean
  - Resp_max

- Variables prefixed by the previous day are prefixed by asterisks.*