Introduction to Modeling

6.872/HST950
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
  \[ P(d | h_3) = 0.5^{10} \approx 0.001 \]
h₁: 100% cherry
h₂: 75% cherry, 25% lime
h₃: 50% cherry, 50% lime
h₄: 25% cherry, 75% lime
h₅: 100% lime

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
Observations

- Bayesian approach asks for prior probabilities on hypotheses!
- Natural way to encode bias against complex hypotheses: make their prior probability very low
- Choosing $h_{\text{MAP}}$ 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
- Regularization is similar to 2nd term—penalty for complexity
- Assuming uniform priors on hypotheses makes MAP yield $h_{\text{ML}}$, the maximum likelihood hypothesis, which maximizes $P(h_i|d) = \alpha P(d|h_i)$
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 \( \beta_j \) to minimize some function of \( (y_i - \hat{y}_i) \) over all \( y_i \)
  - e.g., mean squared error:
  \[ \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n} \]
Logistic Regression

• Logistic function: \( f(z) = \frac{1}{1 + e^{-z}} \)

\[
y_i = f(z_i) \\
z_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_n x_{n,i}
\]

• 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

<table>
<thead>
<tr>
<th></th>
<th>Test Positive</th>
<th>Test Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease Present</td>
<td></td>
<td></td>
</tr>
<tr>
<td>True Positive</td>
<td>TP+FN</td>
<td></td>
</tr>
<tr>
<td>False Negative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease Absent</td>
<td></td>
<td></td>
</tr>
<tr>
<td>False Positive</td>
<td>FP+TN</td>
<td></td>
</tr>
<tr>
<td>True Negative</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 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><strong>Disease Present</strong></td>
<td>True Positive</td>
<td>False Negative</td>
</tr>
<tr>
<td><strong>Disease Absent</strong></td>
<td>False Positive</td>
<td>True Negative</td>
</tr>
<tr>
<td></td>
<td>TP+FP</td>
<td>FN+TN</td>
</tr>
</tbody>
</table>

**Sensitivity (true positive rate):** $\frac{TP}{TP+FN}$

*False negative rate:*$ 1 - Sensitivity = $\frac{FN}{TP+FN}$

**Specificity (true negative rate):** $\frac{TN}{FP+TN}$

*False positive rate:* $1 - $Specificity$ = \frac{FP}{FP+TN}$

**Positive Predictive Value (PPV):** $\frac{TP}{TP+FP}$

**Negative Predictive Value (NPV):** $\frac{TN}{FN+TN}$
Test Thresholds

- FN
- FP

T
Test Thresholds Change Trade-off between Sensitivity and Specificity

- FN
- FP
Receiver Operator Characteristic (ROC) Curve

TPR (sensitivity)

FPR (1-specificity)
What makes a better test?

TPR (sensitivity)

superb

OK

worthless

FPR (1-specificity)
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 test cups
    - Calculate probability that her answers could have occurred by chance guessing; if small, she “wins”
  - ... Fisher’s Exact Test

- Significance testing
  - Reject the null hypothesis (that it happened by chance) if its probability is $< 0.1, 0.05, 0.01, 0.001, ..., 0.000001, ...$
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!*
Overfitting
- bias, variance, noise
- $O = \text{optimal possible model over all possible learners}$
- $L = \text{best model learnable by this learner}$
- $A = \text{actual model learned}$
- $\text{Bias} = O - L$ (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, $U_i$, children, $Y_i$, and children’s parents, $Z_i$. 
Using MIMIC data to build predictive models

- Mortality
  - Comparison to SAPS II
  - Daily Acuity Scores
  - Real-time Acuity Scores
- Other outcomes
  - Good
    - Weaning from Ventilator
    - Weaning from Intra-Aortic Balloon Pump
    - Weaning from Vasopressors
  - Bad
    - Septic shock
    - Hypotension
    - Acute kidney injury

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!

Descriptive look

Figure 3-2: Histograms for demographic information

**Patient Length of Stay**
- **n**: 13923
- **mean**: 5.02
- **median**: 2.67
- **std dev**: 8.1

**Patient Age**
- **n**: 13923
- **mean**: 63.5
- **median**: 65
- **std dev**: 17

**Histogram of Blood Urea Nitrogen (BUN)**
- **n**: 2448468
- **mean**: 30.8
- **median**: 24.3
- **std dev**: 24.3

**Histogram of White Blood Cell Counts**
- **n**: 2448468
- **mean**: 12.9
- **median**: 11.7

**Patient Admit Weight**
- **n**: 13923
- **mean**: 81
- **median**: 75
- **std dev**: 27.2

**Sex of Patients**
- **Female**: n 13923
- **Male**: n 2448468
- **mean**: 81
- **median**: 78.5
- **std dev**: 21.8

**Histogram of Potassium**
- **n**: 2448468
- **mean**: 4.07
- **median**: 4
- **std dev**: 0.534

**Histogram of Sodium**
- **n**: 2448468
- **mean**: 139
- **median**: 139
- **std dev**: 4.76

**Service types for patients**
- **n**: 13923
- **NSICU**: n 325
- **MSICU**: n 396
- **CSRU**: n 83
- **mean**: 24.6
- **median**: 24
- **std dev**: 4.72

**ICD9 Chronic Illnesses**
- **n**: 13923
- **Yes**: n 0
- **No**: n 13923
- **mean**: 24.6
- **median**: 24
- **std dev**: 4.72

Outcomes

Table 3.15: Preprocessed Data

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Patients</td>
<td>10,066</td>
</tr>
<tr>
<td>Number of Rows</td>
<td>1,044,982</td>
</tr>
<tr>
<td>Number of Features</td>
<td>438</td>
</tr>
</tbody>
</table>

### SAPS II

#### 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 positive airway pressure)</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 Score&lt;sup&gt;a&lt;/sup&gt;</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>

<sup>a</sup>If the patient is sedated, the estimated GCS prior to sedation

Training models—5-fold cross validation

Many univariate analyses

<table>
<thead>
<tr>
<th>Obs</th>
<th>Max Deriv Model L.R.</th>
<th>d.f.</th>
<th>P</th>
<th>C</th>
<th>Dxy</th>
</tr>
</thead>
<tbody>
<tr>
<td>20172</td>
<td>1e-09</td>
<td>5415.11</td>
<td>30</td>
<td>0</td>
<td>0.893</td>
</tr>
<tr>
<td>Gamma</td>
<td>Tau-a</td>
<td>0.787</td>
<td>0.176</td>
<td>0.439</td>
<td>0.076</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coef</th>
<th>S.E.</th>
<th>Wald Z P</th>
</tr>
</thead>
<tbody>
<tr>
<td>INR_mean_i</td>
<td>-1.795e+00</td>
<td>1.423e-01</td>
</tr>
<tr>
<td>GCS_max_sq</td>
<td>-7.485e-03</td>
<td>6.000e-04</td>
</tr>
<tr>
<td>OutputB_60_mean_sqrt</td>
<td>-6.561e-02</td>
<td>6.885e-03</td>
</tr>
<tr>
<td>pacemkr_max</td>
<td>-1.084e+00</td>
<td>1.183e-01</td>
</tr>
<tr>
<td>svCSRU_max</td>
<td>-9.516e-01</td>
<td>1.208e-01</td>
</tr>
<tr>
<td>GCSrdv_mean</td>
<td>-1.138e-01</td>
<td>1.528e-02</td>
</tr>
<tr>
<td>pressD01_mean_am</td>
<td>-2.774e+00</td>
<td>3.893e-01</td>
</tr>
<tr>
<td>Platelets_Slope_1680_min</td>
<td>-5.493e+00</td>
<td>8.615e-01</td>
</tr>
<tr>
<td>pressD01_sd_sq</td>
<td>-5.085e+00</td>
<td>8.678e-01</td>
</tr>
<tr>
<td>sedatives_mean_sq</td>
<td>-4.375e+00</td>
<td>8.455e-02</td>
</tr>
<tr>
<td>Bal24_max</td>
<td>-4.493e-05</td>
<td>1.222e-05</td>
</tr>
<tr>
<td>CV_HRRng_max</td>
<td>-3.267e-03</td>
<td>1.083e-03</td>
</tr>
<tr>
<td>Intercept</td>
<td>4.292e+00</td>
<td>4.085e-01</td>
</tr>
<tr>
<td>Milrinone_perKg_min_sq</td>
<td>3.523e+00</td>
<td>1.113e+00</td>
</tr>
<tr>
<td>LOSBal_max</td>
<td>2.247e-05</td>
<td>5.703e-06</td>
</tr>
<tr>
<td>hrnVA_max</td>
<td>3.410e-01</td>
<td>6.767e-02</td>
</tr>
<tr>
<td>MBPm.pr_min_am</td>
<td>1.904e+00</td>
<td>3.711e-01</td>
</tr>
<tr>
<td>Mg_min_sq</td>
<td>1.067e-01</td>
<td>1.798e-02</td>
</tr>
<tr>
<td>beta.Blocking_agent_mean_lam</td>
<td>2.418e-01</td>
<td>3.955e-02</td>
</tr>
<tr>
<td>Na_mean_am</td>
<td>5.214e-02</td>
<td>8.415e-03</td>
</tr>
<tr>
<td>mechVent_mean_sq</td>
<td>7.183e-01</td>
<td>1.047e-01</td>
</tr>
<tr>
<td>RESP_mean_sq</td>
<td>9.226e-04</td>
<td>1.293e-04</td>
</tr>
<tr>
<td>Platelets_mean_i</td>
<td>2.512e+01</td>
<td>3.512e+00</td>
</tr>
<tr>
<td>Lasix_max_lam</td>
<td>2.550e-01</td>
<td>3.457e-02</td>
</tr>
<tr>
<td>CO2_mean_i</td>
<td>2.038e+01</td>
<td>2.741e+00</td>
</tr>
<tr>
<td>jaudiceSkin_mean_la</td>
<td>1.523e-01</td>
<td>2.014e-02</td>
</tr>
<tr>
<td>hospTime_min_sqrt</td>
<td>6.860e-03</td>
<td>7.939e-04</td>
</tr>
<tr>
<td>pressorSum_std_mean_sqrt</td>
<td>7.758e+01</td>
<td>7.225e+02</td>
</tr>
<tr>
<td>Sp02.oor30_t_mean_sqrt</td>
<td>4.929e-01</td>
<td>4.095e-02</td>
</tr>
<tr>
<td>BUNtoCr_min_sqrt</td>
<td>2.867e-01</td>
<td>2.323e-02</td>
</tr>
<tr>
<td>Age_min_q</td>
<td>2.258e-04</td>
<td>1.450e-05</td>
</tr>
</tbody>
</table>

Evaluating the models

**SDAS: All Days**

<table>
<thead>
<tr>
<th>Day</th>
<th>AUC</th>
<th>n</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDAS</td>
<td>0.898</td>
<td>20130</td>
<td>2758</td>
</tr>
</tbody>
</table>

**Devel**

SDAS

- Intercept: 0.005
- Slope: 0.373
- Emax: 0.075

**Validation**

SDAS

- Intercept: 0.006
- Slope: 0.373
- Emax: 0.075
Selected features for each day of ICU stay

DAS model (day n)
