Today’s Quiz

Consider the following semi-supervised dataset.

Consider modeling each class using the Naïve Bayes assumption, with multinomial variables, \( x_i \). Solve using EM, initializing all unlabeled points (" * ") as "+". Give the prediction that would be made at the point marked " ? ".
Today: Missing data

• What's missing?
  
  – Labels
    => Use unsupervised learning
  
  – A subset of observable features, in some of the data examples
    => Today's lecture

• Today's lecture is not a comprehensive treatment of the topic, but rather a case study based on a recent research project:
  

A case study: CATIE study

• CATIE = Clinical Antipsychotic Trial of Intervention and Effectiveness.
  
  – 18 months, 1460 patients with schizophrenia.

• Data collected in a Sequential Multiple Assignment Randomized Trial (SMART).
  
  – Each patient is repeatedly randomized over time.
  – Each randomization occurs at a critical decision point (e.g. milestone in the disease process).
  – Timing and number of randomizations may vary across patients and depend on evolving patient-specific information.
Performance measures

- **Primary outcome:**
  - Minimize “all-cause” treatment discontinuation (incl. efficacy, safety, tolerability).

- **Secondary outcomes:**
  - Symptoms, side effects, vocational and neurocognitive functioning, quality of life, caregiver burden, cost-effectiveness.

- **Scientific goal:** Find the sequence of treatments that produces the best performance according to these outcomes.
### List of variables collected during CATIE

**Variables with no missing information:**

- Time independent variables.
  - Age (cont), Sex (dich), Race (cat), Tardive dyskinesia status at baseline (dich), Marital status at baseline (dich),Patient education (cat), Hospitalization history in 3 months prior to CATIE (dich), Clinical setting in which patient received CATIE treatment (cat), Treatment prior to CATIE enrollment (cat), stage 1 randomized treatment assignment (cat)

**Variables with missing information:**

- Time independent variables.
  - Employment status at baseline (cat), Years since first prescribed anti-psychotic medication at baseline (cont), Neurocognitive composite score at baseline (cont)

- Variables recorded at all months 1-18 and at end-of-stage visits:
  - Adherence measured by the proportion of capsules taken since last visit (cont)

- Variables recorded at months 0, 1, 3, 6, 9, 12, 15, 18 and at end-of-stage visits:
  - Body mass index (cont), Clinical drug use scale (cat), Clinical alcohol use scale (cat), Clinical Global Impressions of Severity of illness score (cat), Positive and Negative Syndrome Scale (cont), Calgary Depression Score (cont), Simpson-Angus EPS mean scale (cont), Barnes Akathisia scale (cont), Total movement severity score (cont)

- Variables recorded at months 0, 6, 12, 18 and at end-of-stage visits:
  - Quality of Life total score (cont), SF-12 Mental health summary (cont), SF-12 Physical health summary (cont), Illicit drug use (dich)

- Variables recorded only at end-of-stage visits:
  - Reason for discontinuing treatment (cat), Stage 2 randomization arm (dich, when applicable), Stage 2 treatment (cat, when applicable)

### Artificial CATIE dataset

<table>
<thead>
<tr>
<th>(G_0)</th>
<th>(W_0)</th>
<th>(P_1)</th>
<th>(A_1, 4)</th>
<th>(W_1)</th>
<th>(P_1)</th>
<th>(C_1)</th>
<th>(A_2)</th>
<th>(P_2)</th>
<th>(W_2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>31.8</td>
<td>103</td>
<td>Perphenazine</td>
<td>23.4</td>
<td>77</td>
<td>SWITCHED</td>
<td>Ziprasidone</td>
<td>86</td>
<td>26.9</td>
</tr>
<tr>
<td>Male</td>
<td>29.4</td>
<td>108</td>
<td>Risperidone</td>
<td>18.2</td>
<td>102</td>
<td>STAYED</td>
<td>NA</td>
<td>88</td>
<td>19</td>
</tr>
<tr>
<td>Male</td>
<td>32.6</td>
<td>63</td>
<td>Olanzapine</td>
<td>35.2</td>
<td>stayed</td>
<td>NA</td>
<td>85</td>
<td>38.2</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>102</td>
<td>102</td>
<td>Quetiapine</td>
<td>34.6</td>
<td>99</td>
<td>SWITCHED</td>
<td>Olanzapine</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>20.8</td>
<td>96</td>
<td>Risperidone</td>
<td>20.8</td>
<td>SWITCHED</td>
<td>Olanzapine</td>
<td>71</td>
<td>31.6</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>38.1</td>
<td>86</td>
<td>Perphenazine</td>
<td>28.7</td>
<td>75</td>
<td>STAYED</td>
<td>NA</td>
<td>88</td>
<td>19</td>
</tr>
<tr>
<td>Female</td>
<td>31.1</td>
<td>80</td>
<td>Risperidone</td>
<td>22.8</td>
<td>89</td>
<td>SWITCHED</td>
<td>Clozapine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>31.6</td>
<td>71</td>
<td>Olanzapine</td>
<td>21.1</td>
<td>STAYED</td>
<td>NA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>25.1</td>
<td>102</td>
<td>Perphenazine</td>
<td>37.9</td>
<td>74</td>
<td>NA</td>
<td>88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>37.9</td>
<td>64</td>
<td>Olanzapine</td>
<td>28.7</td>
<td>91</td>
<td>Risperidone</td>
<td>71</td>
<td>31.6</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>28.7</td>
<td>65</td>
<td>Perphenazine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(W = \text{Body-Mass Index} \)

\(P = \text{PANSS score (measure of symptom intensity)} \)

\(A = \text{Treatment assigned} \)
Missing data in CATIE

- **High study attrition**: only 705 of 1460 stayed for full 18 months; 509 dropped out before entering stage 2.
  - High attrition is not unusual for studies of antipsychotics.

- Majority of missing data (78.1%) was due to attrition.

- We observe a nearly monotone missing data pattern.
  - Monotone: missing data at time $t$ -> missing data at time $t+1$.

- Distribution of most variables appears similar for participants that completed study and those that dropped out.

---

**Figure 1.** Bar plots showing the amount of missing data in the CATIE study. The total height of the bar displays the absolute number of people who have missing (a) PANSS, (b) BMI, and (c) adherence, as measured by pill count, at each of the monthly visits at which the scheduled variable was collected. The dark grey area represents individuals with missing values because they have dropped out of the study prior to that month. The unshaded area is the amount of item missingness in each variable.

- Trend in the amount of missing data over time and proportion of missing data due to dropout are similar for many variables.
Missing data challenges in SMART studies

1. Transition time from one stage to the next depends on patient’s outcome.

2. Some outcome variables are collected at irregularly spaced times (e.g. end of stage), whereas others are collected during regular visits.

3. Some individuals move to follow-up before entering a stage, so we don’t know what treatment they would have been assigned to in the next stage.

4. Some individuals move to follow-up prior to leaving a stage so end-of-stage variables are not collected (incl. randomization prob for next stage.)

5. Observation of some variables depend on patient characteristics and history, resulting in structural missingness.

Types of missing data

• Missing Completely at Random (MCAR)
  – A feature is missing at random, independent of the observed features or the output.

• Missing at Random (MAR)
  – The missing value can depend on other observed variables, but not on the value of the missing feature itself.

• Not Missing at Random (NMAR)
  – The missing value may depend on unobserved variables.

• In general: Hard to detect which case we are dealing with!
Strategies for missing data

- **Deletion methods** => Remove cases (examples) from dataset
  - Listwise deletion
  - Pairwise deletion

- **Substitution methods** (single imputation) => Fill-in missing data
  - Mean/mode substitution
  - Variable control
  - Regression imputation

- **Model-based methods** => Fill in missing data by building model
  - Generative approach of the data
  - Multiple imputation

Listwise deletion (Complete case analysis)

- Only use complete data points.
- Easy to implement!

<table>
<thead>
<tr>
<th>G0</th>
<th>W0</th>
<th>P0</th>
<th>A1</th>
<th>W1</th>
<th>P1</th>
<th>C1</th>
<th>A2</th>
<th>P2</th>
<th>W2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>31.8</td>
<td>103</td>
<td>Perphenazine</td>
<td>23.4</td>
<td>77</td>
<td>SWITCHED</td>
<td>Ziprasidone</td>
<td>86</td>
<td>26.9</td>
</tr>
<tr>
<td>Male</td>
<td>29.4</td>
<td>108</td>
<td>Risperidone</td>
<td>18.2</td>
<td>102</td>
<td>STAYED</td>
<td>NA</td>
<td>88</td>
<td>19</td>
</tr>
<tr>
<td>Male</td>
<td>32.6</td>
<td>63</td>
<td>Olanzapine</td>
<td>35.2</td>
<td></td>
<td></td>
<td></td>
<td>85</td>
<td>38.2</td>
</tr>
<tr>
<td>Female</td>
<td>102</td>
<td>Quetiapine</td>
<td>34.6</td>
<td>99</td>
<td>SWITCHED</td>
<td>Olanzapine</td>
<td>77</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>31.1</td>
<td>80</td>
<td>Risperidone</td>
<td>22.8</td>
<td>89</td>
<td>SWITCHED</td>
<td>Clozapine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>31.6</td>
<td>71</td>
<td>Olanzapine</td>
<td>21.1</td>
<td></td>
<td>STAYED</td>
<td>NA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>38.1</td>
<td>86</td>
<td>Perphenazine</td>
<td>28.7</td>
<td>75</td>
<td>STAYED</td>
<td>NA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>31.1</td>
<td>80</td>
<td>Risperidone</td>
<td>22.8</td>
<td>89</td>
<td>SWITCHED</td>
<td>Clozapine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>31.6</td>
<td>71</td>
<td>Olanzapine</td>
<td>21.1</td>
<td></td>
<td>STAYED</td>
<td>NA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>25.1</td>
<td>64</td>
<td>Olanzapine</td>
<td>19.7</td>
<td>74</td>
<td>STAYED</td>
<td>NA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>37.9</td>
<td>64</td>
<td>Olanzapine</td>
<td>36</td>
<td></td>
<td>STAYED</td>
<td>NA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>28.7</td>
<td>91</td>
<td>Risperidone</td>
<td>22.2</td>
<td>89</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>37.8</td>
<td>65</td>
<td>Perphenazine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Pairwise deletion (Available case analysis)

- Use all cases in which the variables of interest are present.
  - E.g. Decision tree: evaluate test on \( x_i \), using examples with that var.
- Uses as much information as possible.
  - Difficult to analyze since using different feature vectors. Bias if not MCAR.

<table>
<thead>
<tr>
<th>G₀</th>
<th>W₀</th>
<th>P₀</th>
<th>A₁</th>
<th>W₁</th>
<th>P₁</th>
<th>C₁</th>
<th>A₂</th>
<th>P₂</th>
<th>W₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>31.8</td>
<td>103</td>
<td>Perphenazine</td>
<td>23.4</td>
<td>77</td>
<td>SWITCHED</td>
<td>Ziprasidone</td>
<td>86</td>
<td>26.9</td>
</tr>
<tr>
<td>Male</td>
<td>29.4</td>
<td>108</td>
<td>Risperidone</td>
<td>18.2</td>
<td>102</td>
<td>STAYED</td>
<td>NA</td>
<td>88</td>
<td>19</td>
</tr>
<tr>
<td>Male</td>
<td>32.6</td>
<td>63</td>
<td>Olanzapine</td>
<td>35.2</td>
<td>—</td>
<td>STAYED</td>
<td>NA</td>
<td>85</td>
<td>38.2</td>
</tr>
<tr>
<td>Female</td>
<td>—</td>
<td>102</td>
<td>Quetiapine</td>
<td>34.6</td>
<td>99</td>
<td>SWITCHED</td>
<td>Olanzapine</td>
<td>77</td>
<td>—</td>
</tr>
<tr>
<td>Female</td>
<td>—</td>
<td>—</td>
<td>Risperidone</td>
<td>20.8</td>
<td>96</td>
<td>SWITCHED</td>
<td>Olanzapine</td>
<td>71</td>
<td>31.6</td>
</tr>
<tr>
<td>Male</td>
<td>38.1</td>
<td>86</td>
<td>Perphenazine</td>
<td>28.7</td>
<td>75</td>
<td>STAYED</td>
<td>NA</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Female</td>
<td>31.1</td>
<td>80</td>
<td>Risperidone</td>
<td>22.8</td>
<td>89</td>
<td>SWITCHED</td>
<td>Clozapine</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Female</td>
<td>31.6</td>
<td>71</td>
<td>Olanzapine</td>
<td>21.1</td>
<td>—</td>
<td>STAYED</td>
<td>NA</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Male</td>
<td>25.1</td>
<td>64</td>
<td>Olanzapine</td>
<td>19.7</td>
<td>74</td>
<td>STAYED</td>
<td>NA</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Male</td>
<td>37.9</td>
<td>64</td>
<td>Olanzapine</td>
<td>36</td>
<td>—</td>
<td>STAYED</td>
<td>NA</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Female</td>
<td>28.7</td>
<td>91</td>
<td>Risperidone</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Male</td>
<td>37.8</td>
<td>65</td>
<td>Perphenazine</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Strategies for missing data

- **Deletion methods** => Remove cases (examples) from dataset
  - Listwise deletion
  - Pairwise deletion

- **Substitution methods** (single imputation) => Fill-in missing data
  - Mean/mode substitution
  - Variable control
  - Regression imputation

- **Model-based methods** => Fill in missing data by building model
  - Generative approach
  - Multiple imputation
Mean / Mode substitution

- Replace missing value with sample mean or mode.
- Train learner as if all complete cases.

Advantages:
- Easy to implement!

Disadvantages:
- Bias unless MCAR.
- Reduces variability.
- Weakens covariance and correlation estimates in the data because it ignores relationship between variables.

Variable control

- **Add a binary indicator variable** (1 = value is missing; 0 = value is observed) to model missingness for each variable.
- Fill-in missing values using a constant (e.g. the sample mean).
- Train learner as in complete case, including indicator variables.

Advantage:
- Uses all available information about missing observation.

Disadvantage:
- Results in biased estimates, unless MCAR.
Regression imputation

• Replace missing values with predicted value from a regression equation.

• Advantage:
  – Uses information from observed data.

• Disadvantage;
  – Overestimates model fit and correlation estimates. Weakens variance.

Strategies for missing data

• **Deletion methods** => Remove cases (examples) from dataset
  – Listwise deletion
  – Pairwise deletion

• **Substitution methods** (single imputation) => Fill-in missing data
  – Mean/mode substitution
  – Variable control
  – Regression imputation

• **Model-based methods** => Fill in missing data by building model of the data
  – Generative approach
  – Multiple imputation
Generative approach

• Assume a joint probabilistic model for the data.
• Estimate the maximum likelihood setting for the missing data using Expectation-Maximization.

• Advantages:
  – Uses full information to calculate likelihood.
  – Unbiased parameter estimation for MCAR/MAR cases.
• Disadvantages:
  – Converges to a local minima.

Multiple imputation

• Imputation: Data is “filled in” with predicted values from a trained regression model.
• Need a good regression model to get good imputations.

• Repeat imputation \( k \) times, producing \( k \) separate datasets
• Train predictor for each imputed (complete) dataset and merge results into one estimate (e.g. majority voting).

• This is the approach we implemented for CATIE.
Multiple imputation in CATIE

- Fit a (separate) conditional model for each variable.
- Algorithm:
  - Let \( v_{t,p}, \ldots, v_{t,J} \) denote the variables collected at time \( t \).
  - Order these variables according to amount of missingness.
  - Let \( D_{t-1} = \{ v_{0p}, v_{t,1p}, \ldots, v_{t,Jt} \} \).
  - Estimate the joint posterior predictive distribution of the missing observations given the observed variables:
    \[
    \int \cdots \int \prod_{t=1}^{T} \prod_{j=1}^{J} f(v_{t,j}|D_{t-1}, \theta_{t,j}) \pi(\theta_{t,j}|D_{t-1}, v_{t,j,\text{obs}}, \theta_{1,1}, \ldots, \theta_{t-1,1}, \ldots, \theta_{t,j-1}) d\theta_{t,j}
    \]
  - First term is the conditional on the current variable. Second term is the prior on the parameters of the distribution (\( \theta \)).
  - The posterior is estimated by sampling, time step by time step.

Multiple imputation in CATIE

- Using separate models for each variable is computationally advantageous (compared to full joint distribution over all variables.)
- But can lead to unrealistic fluctuations in some variables over time.
- Challenge: impose smoothness constraint (over time) on some variables.
- Solution: Use spline regression to enforce smoothness over time on the conditional mean.
Multiple imputation in CATIE

Overall imputation strategy:
1. Impute baseline variables (only 3% of data is missing).
2. Impute stage transition times. Use single imputation for this.
3. Impute end-of-stage variables.
   - Pool data over multiple time-windows (months) to get better estimates.
4. Impute randomly assigned treatment (especially for stage 2).
5. Impute additional missing time-varying information.

Imputed vs Observed PANSS scores

Figure 1. Histograms for (a) observed month of entry into stage 2 of CATIE, (b) singly imputed stage 2 transition time for those CATIE participants (c) for individuals initially assigned to perphenazine observed and singly imputed stage 2 transition times (d) multiply imputed stage 2 transition time for individuals assigned to perphenazine in stage 1.

Figure 2. QQ-plots of imputed versus observed PANSS scores measured at all months in which PANSS was scheduled to be collected and all end-of-stage PANSS scores. The missing data distribution contains the imputed values from twenty-five imputations (and none of the observed values).

In this case, we notice non-trivial differences between the observed and imputed distributions. In particular, many more people have lower adherence in the imputed data than in the observed. While this certainly raises a red flag, it does not necessarily mean that the imputations are not valid. Recall that CATIE participants were allowed to discontinue treatment, or drop out of the study, for any reason including adherence. In fact, this aspect of the CATIE protocol resulted in many non-adherent patients switching into the next treatment stage, or dropping out of the study, rather than remaining non-adherent to their current treatment. This resulted in very high recorded adherence rates throughout the CATIE study, with the median recorded adherence at each month ranging from 75% to 100%. CATIE participants with adherence below 50% at a monthly visit compared to those who had adherence higher than 50% had a log odds ratio of dropping out of the study before the next monthly visit of 1.82, with a standard error...
Imputed vs Observed BMI values

Imputed vs Observed Adherence

Imputation seems less accurate:
lower adherence in imputed data.
Maybe this is ok:
Low adherence is common cause
of drop-out, so not surprising to
see many cases that needed
imputation have low adherence.
CATIE analysis with imputed data

Table 2. Estimated mean PANSS score over the 18 months of the CATIE study for each of the treatment regimes and 95% confidence intervals. The columns entitled Complete Case report the number of people (N) contributing information to estimating the mean response for each regime, the estimated mean response and corresponding 95% CI. The columns entitled Multiple Imputation report the number of people (N) averaged over 25 imputations contributing information to estimating the mean response for each regime as well as the estimated mean response and 95% CI.

<table>
<thead>
<tr>
<th>Treatment Regimes</th>
<th>Complete Case</th>
<th>Multiple Imputation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Mean [95% CI]</td>
</tr>
<tr>
<td>Olanzapine,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>If fail to respond, then</td>
<td>89</td>
<td>62.63 [60.22, 65.04]</td>
</tr>
<tr>
<td>Risperidone</td>
<td>92</td>
<td>64.14 [60.98, 67.30]</td>
</tr>
<tr>
<td>If due to efficacy clozapine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>If fail to respond, then</td>
<td>97</td>
<td>62.73 [60.02, 64.44]</td>
</tr>
<tr>
<td>Quetiapine,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>If fail to respond, then</td>
<td>47</td>
<td>63.88 [59.65, 68.11]</td>
</tr>
<tr>
<td>Risperidone</td>
<td>48</td>
<td>65.89 [62.30, 69.47]</td>
</tr>
<tr>
<td>If due to efficacy clozapine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>If fail to respond, then</td>
<td>52</td>
<td>65.67 [61.54, 69.80]</td>
</tr>
<tr>
<td>Risperidone</td>
<td></td>
<td></td>
</tr>
<tr>
<td>If fail to respond, then</td>
<td>74</td>
<td>65.93 [61.91, 69.94]</td>
</tr>
<tr>
<td>Quetiapine,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>If fail to respond, then</td>
<td>71</td>
<td>68.52 [64.80, 72.24]</td>
</tr>
<tr>
<td>Olanzapine,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>If due to efficacy clozapine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>If due to tolerance ziprasidone</td>
<td>71</td>
<td>66.71 [63.06, 70.35]</td>
</tr>
</tbody>
</table>

Final comments

- Missing data can cause significant bias in analysis.
- Many methods for handling missing data; in general, need to understand your data and missingness pattern to figure out what technique is appropriate.
- EM algorithm can be used to estimate parameters of generative model and fill-in missing data.
- Multiple imputation is a successful method for cases with structural missingness, but requires significant modeling effort.
Notes

• Final project:
  – Submit slides (.pdf only) for 3min talk before 10am on Wednesday Dec.3 by email (jpineau@cs.mcgill.ca).
  – Submit project report by Dec.11, 11:59pm, on CMT (Track #4).

• Complete online course evaluations on Minerva!

• Midterms are available for viewing during regular office hours (Tuesdays, 12:00-1:30pm).