CS340

Bayesian concept learning cont'd

Kevin Murphy
Healthy levels game

"healthy levels"
Hypothesis space

\[ h = (\ell_1, \ell_2, s_1, s_2) \]

Healthy levels of insulin/cholesterol must lie between a minimum and maximum. Healthy levels of a chemical presumably lie between zero and a maximum.
Likelihood (strong sampling)

- \( p(X|h) = \frac{1}{|h|^n} \) if all \( x_i \in h \), where \( |h| = s_1 \times s_2 \)
- \( p(X|h) = 0 \) if any \( x_i \) outside \( h \)
Prior $p(h)$

- Use uninformative, but location and scale-invariant, prior (Jeffrey’s principle)

$$p(h) \propto \frac{1}{s_1 s_2}$$

This also happens to be conjugate to $p(X|h)$.

- We will explain this later...
Posterior predictive

\[ p(y \in C | X) = \int_{h \in H} p(y \in C | h)p(h | X) dh \]

Since the hypothesis space is continuous, we must use an integral instead of a sum...
Insert hairy math

\[ l - \sigma \leq -\tau, \text{ where } \sigma \text{ is size of the rectangle. Hence} \]

\[
p(X) = \int_{H_{\Sigma}} \frac{p(h)}{|h|} \, dh \quad \text{(1.34)}
\]

\[
= \int_{H_{\Sigma}} \left( -e^{-\frac{1}{\sigma^2}} \right) \, dh
\]

\[
= \int_{H_{\Sigma}} \left( -e^{-\frac{1}{\sigma^2}} \right) \, dh 
= \int_{\text{area}} \left( -e^{-\frac{1}{\sigma^2}} \right) 
= \int_{\text{area}} \left( -e^{-\frac{1}{\sigma^2}} \right) 
= \int_{\text{area}} \left( -e^{-\frac{1}{\sigma^2}} \right) 
= \int_{\text{area}} \left( -e^{-\frac{1}{\sigma^2}} \right) 
= \int_{\text{area}} \left( -e^{-\frac{1}{\sigma^2}} \right) 
= \int_{\text{area}} \left( -e^{-\frac{1}{\sigma^2}} \right) 
= \int_{\text{area}} \left( -e^{-\frac{1}{\sigma^2}} \right)
\]

Now, using integration by parts:

\[
l = \int_a^b f(s)g(s) \, ds = [f(s)g(s)]_a^b - \int_a^b f'(s)g(s) \, ds
\]

with the substitutions

\[
f(s) = \hat{z} - \tau \quad \text{(1.40)}
\]

\[
f'(s) = 1 \quad \text{(1.41)}
\]

\[
f''(s) = x^{n-1} \quad \text{(1.42)}
\]

\[
g(s) = \frac{x^n}{n} \quad \text{(1.43)}
\]

we have

\[
p(X) = \left[ \frac{n}{n+1} \right]^{n+1}_{n+1} \int_1^\infty \frac{\tau^{n+1}}{\tau^{n+1}} \, d\tau
\]

\[
= \left[ \frac{n}{n+1} \right]^{n+1}_{n+1} \int_1^\infty \frac{\tau^{n+1}}{\tau^{n+1}} \, d\tau
\]

\[
= \left[ \frac{n}{n+1} \right]^{n+1}_{n+1} \int_1^\infty \frac{\tau^{n+1}}{\tau^{n+1}} \, d\tau
\]

\[
= \left[ \frac{n}{n+1} \right]^{n+1}_{n+1} \int_1^\infty \frac{\tau^{n+1}}{\tau^{n+1}} \, d\tau
\]

\[
= \left[ \frac{n}{n+1} \right]^{n+1}_{n+1} \int_1^\infty \frac{\tau^{n+1}}{\tau^{n+1}} \, d\tau
\]

\[
= \left[ \frac{n}{n+1} \right]^{n+1}_{n+1} \int_1^\infty \frac{\tau^{n+1}}{\tau^{n+1}} \, d\tau
\]

\[
= \left[ \frac{n}{n+1} \right]^{n+1}_{n+1} \int_1^\infty \frac{\tau^{n+1}}{\tau^{n+1}} \, d\tau
\]

\[
= \left[ \frac{n}{n+1} \right]^{n+1}_{n+1} \int_1^\infty \frac{\tau^{n+1}}{\tau^{n+1}} \, d\tau
\]

\[
= \left[ \frac{n}{n+1} \right]^{n+1}_{n+1} \int_1^\infty \frac{\tau^{n+1}}{\tau^{n+1}} \, d\tau
\]

\[
= \left[ \frac{n}{n+1} \right]^{n+1}_{n+1} \int_1^\infty \frac{\tau^{n+1}}{\tau^{n+1}} \, d\tau
\]

\[
= \left[ \frac{n}{n+1} \right]^{n+1}_{n+1} \int_1^\infty \frac{\tau^{n+1}}{\tau^{n+1}} \, d\tau
\]

\[
= \left[ \frac{n}{n+1} \right]^{n+1}_{n+1} \int_1^\infty \frac{\tau^{n+1}}{\tau^{n+1}} \, d\tau
\]

\[
= \left[ \frac{n}{n+1} \right]^{n+1}_{n+1} \int_1^\infty \frac{\tau^{n+1}}{\tau^{n+1}} \, d\tau
\]

\[
= \left[ \frac{n}{n+1} \right]^{n+1}_{n+1} \int_1^\infty \frac{\tau^{n+1}}{\tau^{n+1}} \, d\tau
\]

\[
= \left[ \frac{n}{n+1} \right]^{n+1}_{n+1} \int_1^\infty \frac{\tau^{n+1}}{\tau^{n+1}} \, d\tau
\]

To compute the generalization function, let us suppose \( \hat{y} \) is outside the range spanned by the sample (otherwise the probability of generalization is 1). Without loss of generality assume \( \hat{y} > 0 \). Let \( \delta \) be the distance from \( \hat{y} \) to the closest observed example. Then we can compute the numerator in Equation 1.33 by replacing \( \tau \) with \( \hat{y} + \delta \) in the limits of integration (since we have expanded the range of the data by adding \( \hat{y} \)), yielding

\[
p(y \in \mathcal{C} | X) = \int_{H \in \mathcal{X}} \frac{\hat{y}(h)}{|h|^q} \, dh
\]

\[
= \int_{-\delta}^{\delta} \int_{H \in \mathcal{X}} \frac{\hat{y}(h)}{|h|^q} \, dh
\]

\[
= \frac{1}{n(n-1)q^{n+1}}
\]
And the answer is…

\[ p(y \in C|X) = \left[ \frac{1}{(1 + \tilde{d}_1/r_1)(1 + \tilde{d}_2/r_2)} \right]^{n-1} \]

\( \tilde{d}_i = 0 \) if \( y \in \) range of \( X_i \)

= distance of \( y \) from closest \( X_i \)
Behavior for n=3, 6, 12

The size principle implies the smallest rectangle has highest likelihood, but there are many other consistent rectangles which are only slightly less likely. These get averaged to give a smooth generalization gradient.

As $N \to \infty$, the larger hypotheses become exponentially less likely, so we converge on the ML solution (the most specific/ MIN hypothesis)
Behavior for different shapes

- $n=3$ in both cases, but on right, $r_1 << r_2$, so we generalize more along dimension 2.
- Algebraically, $d_1/r_1$ is big, so $p(y \in C \mid X)$ is small unless $y$ is inside $X$.
- Intuitively, it would be a suspicious coincidence if the rectangle was wide but $r_1$.

\[
p(y \in C \mid X) = \left[ \frac{1}{(1 + \tilde{d}_1/r_1)(1 + \tilde{d}_2/r_2)} \right]^{n-1}
\]
Behavior of max likelihood/ MAP

There is no generalization gradient (a point is either in or out of h). The ML/MAP hyp. is the smallest enclosing rectangle. This is a good approximation to Bayes when N is large.
Weak sampling

• Examples are not sampled from the concept, they are just labeled as consistent or not.

\[ p(X|h) = \begin{cases} 
1 & \text{if } x_1, \ldots, x_n \in h \\
0 & \text{if any } x_i \notin h 
\end{cases} \]
Behavior of weak Bayes

We do not get convergence to the ML hypothesis. If truth is a rectangle, we do not converge to it (not a consistent estimator).
A more realistic example

• A discrete hypothesis space (the number game)
• A continuous hypothesis space (the healthy levels concept)
• Word learning
Here is a pog:

Can you give Mr. Frog all the other pogs?
Hierarchical categories

<table>
<thead>
<tr>
<th></th>
<th>Vegetables</th>
<th>Vehicles</th>
<th>Animals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>subordinate</strong></td>
<td>![image]</td>
<td>![image]</td>
<td>![image]</td>
</tr>
<tr>
<td><strong>basic</strong></td>
<td>![image]</td>
<td>![image]</td>
<td>![image]</td>
</tr>
<tr>
<td><strong>superordinate</strong></td>
<td>![image]</td>
<td>![image]</td>
<td>![image]</td>
</tr>
</tbody>
</table>
Human data

Example sets:

1 subordinate

3 subordinate

3 basic

3 superordinate

Vegetables

Vehicles

Animals

Green peppers? All peppers? All veg?

Generalize up to least common ancestor
Hypothesis space

Structure

Superordinate level

Basic level

Subordinate level

Examples:

Data

'fep'  'fep'  'fep'  'dax'  'zoog'  'gazzer'  ...
Hypothesis space

Derived by applying agglomerative clustering to human similarity matrix
Hierarchical Clustering

- Cluster based on similarities/distances
- Distance measure between instances $\mathbf{x}^r$ and $\mathbf{x}^s$

Minkowski ($L_p$) (Euclidean for $p = 2$)

$$d_m(\mathbf{x}^r, \mathbf{x}^s) = \left( \sum_{j=1}^{d} (x^r_j - x^s_j)^p \right)^{1/p}$$

City-block distance

$$d_{cb}(\mathbf{x}^r, \mathbf{x}^s) = \sum_{j=1}^{d} |x^r_j - x^s_j|$$
Agglomerative Clustering

• Start with $N$ groups each with one instance and merge two closest groups at each iteration

• Distance between two groups $G_i$ and $G_j$:
  – Single-link:
    \[ d(G_i, G_j) = \min_{x^r \in G_i, x^s \in G_j} d(x^r, x^s) \]
  – Complete-link:
    \[ d(G_i, G_j) = \max_{x^r \in G_i, x^s \in G_j} d(x^r, x^s) \]
  – Average-link, centroid
Example: Single-Link Clustering
\[
p(h) = \text{height}(\text{parent}(h)) - \text{height}(h)
\]

\[
p(X|h) = \left[ \frac{1}{\text{height}(h)} \right]^n
\]
Strong Bayes (w/ basic-level bias)

Example sets:

1 subordinate

3 subordinate

3 basic

3 superordinate

Vegetables

Vehicles

Animals

---

[Bar charts showing data and model predictions for different example sets across categories of vegetables, vehicles, and animals.]
Word learning vs healthy levels

- In the word domain, after about $N=3$ we have an "aha" moment (rule-like learning), but for healthy levels, we need a large sample size, because in the former, hypotheses differ dramatically in size, so we rapidly prefer the smallest consistent, whereas latter averages many.
Rules and exemplars in the number game

• Hyp. space is a mixture of sparse (mathematical concepts) and dense (intervals) hypotheses.

• If data supports mathematical rule (eg $X=\{16,8,2,64\}$), we rapidly learn a rule, otherwise (eg $X=\{6,23,19,20\}$) we learn by similarity, and need many examples to get sharp boundary.