CS340

Bayesian concept learning cont'd

Kevin Murphy
Homework 2

• Bring in a paper copy to class on Monday
• If you can’t come to class, ask a friend to bring it, or use dropbox #10.
• If you use the dropbox, please email the TAs to tell them to pick it up.
• No need to use handin anymore!
Summary of the Bayesian approach

1. Constrained hypothesis space $H$
2. Prior $p(h)$
3. Likelihood $p(X|h)$
4. Hypothesis (model) averaging:

$$p(y \in C | X) = \sum_h p(y \in C | h)p(h | X)$$
Maximum likelihood

- ML learning finds the most likely hypothesis and then uses the plug-in principle for prediction.
  $$\hat{h} = \arg \max_h p(X|h)$$
  $$p(y \in C|X) = p(y \in C|\hat{h})$$

- Given $X=\{16\}$, $\hat{h} = "\text{powers of 4}"$, given $X=\{16,8,2,64\}$, $\hat{h} = "\text{powers of 2}"$.

- So predictive distribution gets broader as we get more data, in contrast to bayes.
Maximum likelihood

• As the amount of data goes to $\infty$, ML and Bayes converge to the same solution, since the likelihood overwhelms the prior, since $p(X|h)$ grows with $N$, but $p(h)$ is constant.

• This is not true if we use weak sampling model, $p(X|h) = \delta(X \in H_x)$

• If truth is in the hypothesis class, both methods will find it; thus they are both consistent estimators.
MAP (maximum a posterior) learning

- We find the mode of the posterior, and use it as a plug-in.
  \[ \hat{h} = \arg \max_h p(h|X) = \arg \max_h p(X|h)p(h) \]
  \[ p(y \in C|X) = p(y \in C|\hat{h}) \]

- As \( N \to \infty \), the posterior peaks around the mode, so MAP/ML/Bayes solution converge
  \[ p(y \in C|X) = \sum_{h} p(y \in C|h)p(h|X) \to \sum_{h} p(y \in C|h)\delta(h, \hat{h})) = p(y \in C|\hat{h}) \]

- Cannot explain transition from similarity-based (broad posterior) to rule-based (narrow posterior)
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

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

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

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

\[ p(X) = \int_{0}^{\infty} \frac{p(\theta)}{\theta} d\theta \]

\[ = \int_{0}^{\infty} \int_{0}^{\theta} \frac{p(s)}{s} ds \; d\theta \]

\[ = \int_{0}^{\infty} \left[ \int_{0}^{t} \frac{1}{s^n} ds \right] \theta^{-n} \; d\theta \]

\[ = \int_{0}^{\infty} \frac{1}{\theta^n} \theta^{-n} \; d\theta \]

\[ = \int_{0}^{\infty} \theta^{-2n+1} \; d\theta \]

Now, using integration by parts.

\[ l = \int_{a}^{b} f(x)g(x)dx = [f(x)g(x)]_a^b - \int_{a}^{b} f'(x)g(x)dx \]

with the substitutions

\[ f(x) = \frac{d}{dx} \]

\[ f'(x) = 1 \]

\[ g(x) = x^{-n} \]

we have

\[ p(X) = \left[ \int_{0}^{\infty} \frac{\theta^{-n}}{\theta} d\theta \right] \int_{a}^{b} f'(x)g(x)dx \]

\[ = \left[ \frac{1}{\int_{-\infty}^{\infty} \theta^{-n+1} \; d\theta} \right] \int_{a}^{b} f'(x)g(x)dx \]

\[ = \frac{1}{1} \frac{1}{\int_{a}^{b} f'(x)g(x)dx} \]

\[ = \frac{1}{\int_{a}^{b} f'(x)g(x)dx} \]

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

\[ p(y \in S, X) = \int_{\nu \in \mathbb{R}} \frac{p(\nu)}{\nu} d\nu \]

\[ = \int_{-\infty}^{\infty} \int_{0}^{\infty} \frac{p(\nu)}{\nu} d\nu \; ds \]

\[ = \int_{0}^{\infty} \frac{1}{\nu(n-1)(\nu + d)^{n-1}} \]

1-17
And the answer is...

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

\[ \tilde{d}_i = 0 \text{ if } y \in \text{ range of } X_i \]
\[ = \text{ distance of } y \text{ from closest } X_i \]