CS340

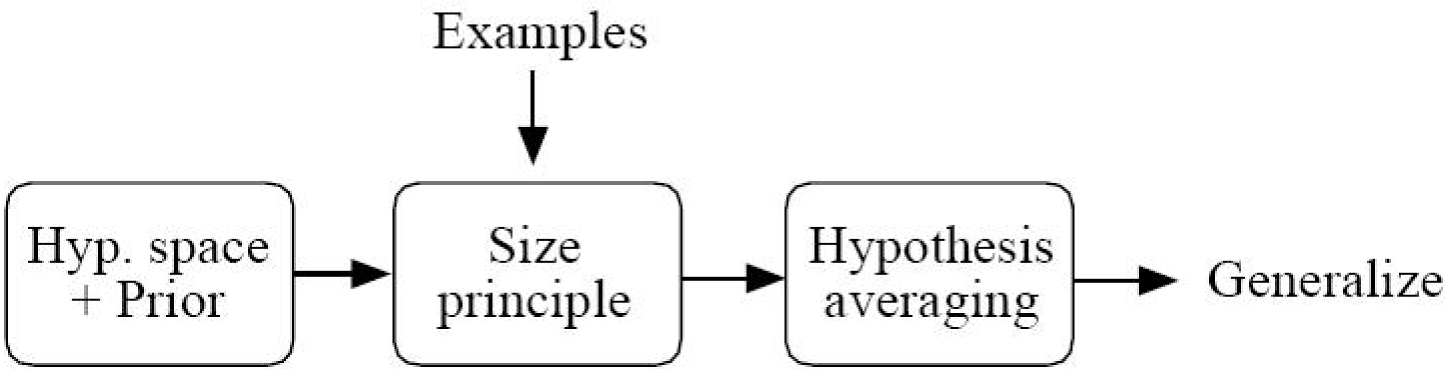
Bayesian concept learning cont'd

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. $\hat{h}$

$$\begin{aligned} \hat{h} &= \arg\max_h p(X|h) \\ p(y \in C|X) &= p(y \in C|\hat{h}) \end{aligned}$$

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

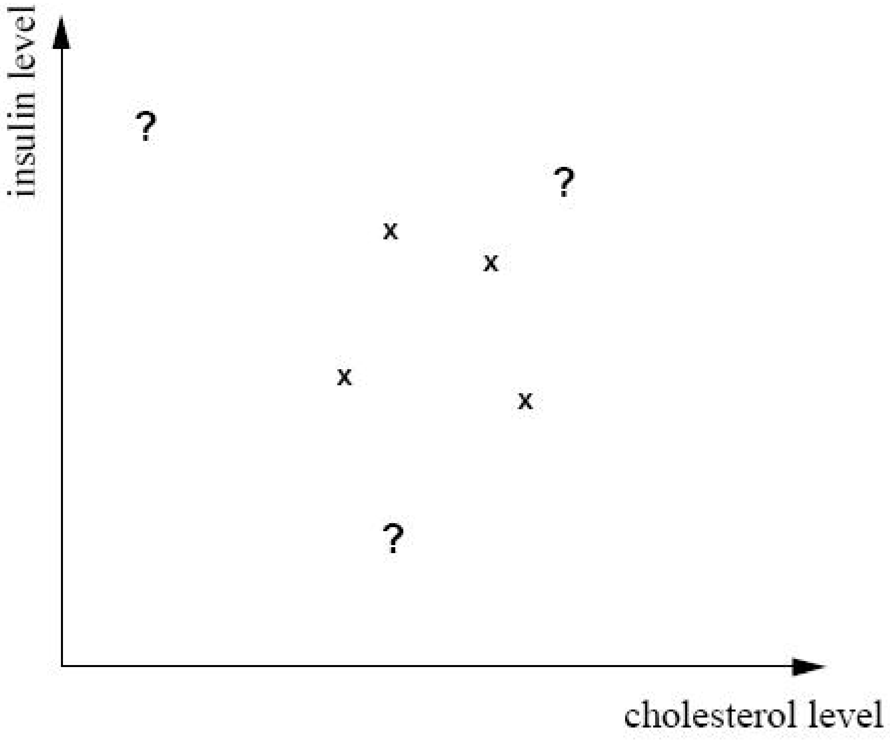
$$\begin{aligned} \hat{h} &= \arg\max_h p(h|X) = \arg\max_h p(X|h)p(h) \\ p(y \in C|X) &= p(y \in C|\hat{h}) \end{aligned}$$

- As N $\rightarrow \infty$, the posterior peaks around the mode, so MAP/ML/Bayes solution converge

$$p(y \in C|X) = \sum_h p(y \in C|h)p(h|X) \rightarrow \sum_h p(y \in C|h)\delta(h, \hat{h})) = p(y \in C|\hat{h})$$

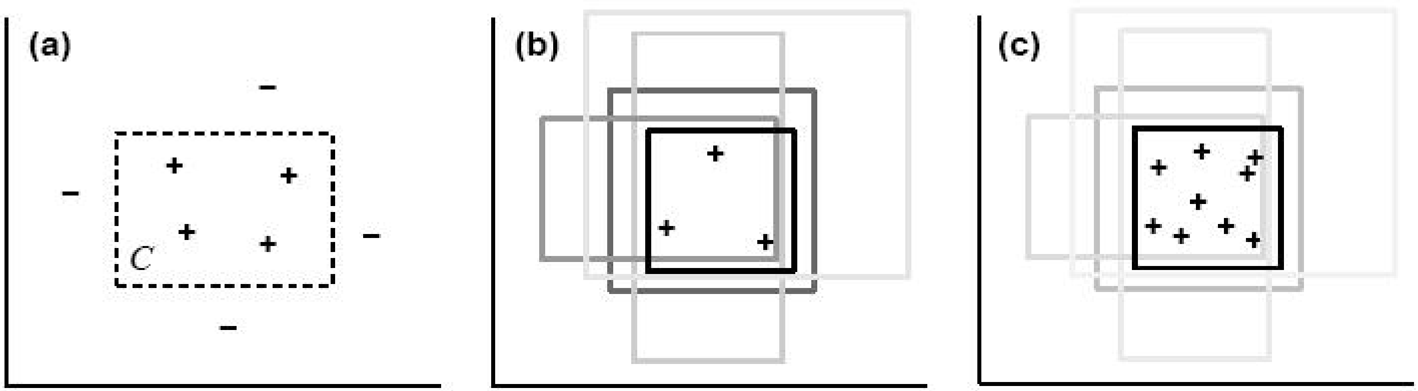
- Cannot explain transition from similarity-based (broad posterior) to rule-based (narrow posterior)

Healthy levels game



"healthy levels"

Hypothesis space



$$h = (\ell_1, \ell_2, s_1, s_2)$$

Healthy levels of insulin/ cholestrol must lie between a minimum and maximum. Healthy levels of a chemical presumably lie between zero and a maximum.

Likelihood (strong sampling)

- $p(X|h) = 1/|h|^n$ 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

$l - s \le -r$, where s is size of the rectangle. Hence

$$
\begin{aligned}
p(X) &= \int_{h \in \mathcal{H}_X} \frac{p(h)}{|h|^n} dh &(1.34)\\
&= \int_{s=r}^{\infty} \int_{l=0}^{s-r} \frac{p(s)}{s^n} dl ds &(1.35)\\
&= \int_{s=r}^{\infty} \left[\int_{l=0}^{s-r} \frac{1}{s^{n+1}} dl \right] ds &(1.36)\\
&= \int_{s=r}^{\infty} \frac{1}{s^{n+1}} [l]_0^{s-r} ds &(1.37)\\
&= \int_{s=r}^{\infty} \frac{s-r}{s^{n+1}} ds &(1.38)
\end{aligned}
$$

Now, using **integration by parts**

$$I = \int_a^b f(x) g'(x) dx = [f(x)g(x)]_a^b - \int_a^b f'(x) g(x) dx \quad (1.39)$$

with the substitutions

$$
\begin{aligned}
f(s) &= s - r &(1.40)\\
f'(s) &= 1 &(1.41)\\
f'(s) &= s^{-n-1} &(1.42)\\
g(s) &= \frac{s^{-n}}{-n} &(1.43)
\end{aligned}
$$

we have

$$
\begin{aligned}
p(X) &= \left[\frac{(s-r)s^{-n}}{-n} \right]_r^{\infty} - \int_r^{\infty} \frac{s^{-n}}{-n} ds &(1.44)\\
&= \left[\frac{s^{-n+1}}{-n} + \frac{rs^{-n}}{n} - \frac{-1}{n} \frac{s^{-n+1}}{-n+1} \right]_r^{\infty} &(1.45)\\
&= \frac{r^{-n+1}}{n} - \frac{rr^{-n}}{n} + \frac{r^{-n+1}}{n(n-1)} &(1.46)\\
&= \frac{1}{nr^{n-1}} - \frac{r}{nr^{n-1}r} + \frac{1}{n(n-1)r^{n-1}} &(1.47)\\
&= \frac{1}{n(n-1)r^{n-1}} &(1.48)
\end{aligned}
$$

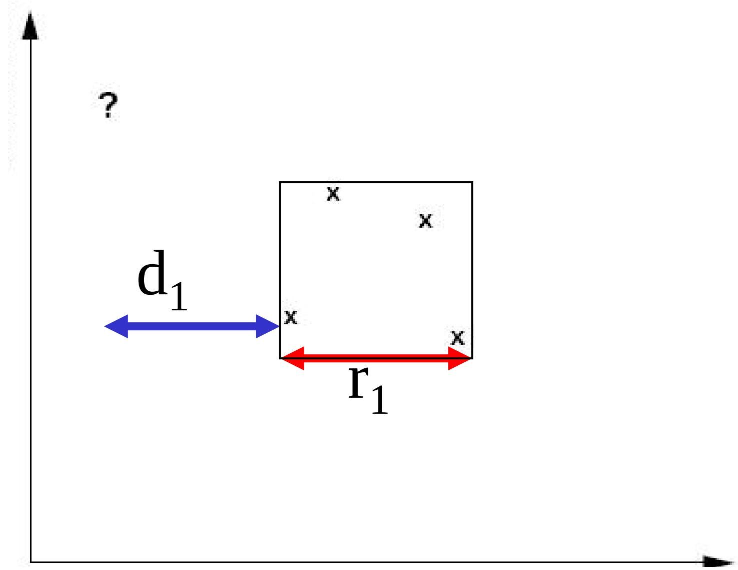
To compute the generalization function, let us suppose y is outside the range spanned by the examples (otherwise the probability of generalization is 1). Without loss of generality assume $y > 0$. Let d be the distance from y to the closest observed example. Then we can compute the numerator in Equation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

$$
\begin{aligned}
p(y \in C, X) &= \int_{h \in \mathcal{H}_{X,y}} \frac{p(h)}{|h|^n} dh &(1.49)\\
&= \int_{r+d}^{\infty} \int_0^{s-(r+d)} \frac{p(s)}{s^n} dl ds &(1.50)\\
&= \frac{1}{n(n-1)(r+d)^{n-1}} &(1.51)
\end{aligned}
$$

1-17

And the answer is…

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



$$\begin{aligned}\tilde{d}_i &= 0 \text{ if } y \in \text{ range of } X_i \\ &= \text{ distance of y from closest } X_i\end{aligned}$$