

Outline

- ◇ Bayesian learning
- ◇ Maximum *a posteriori* and maximum likelihood learning
- ◇ Bayes net learning
 - ML parameter learning with complete data
 - linear regression

Full Bayesian learning

View learning as Bayesian updating of a probability distribution over the **hypothesis space**

H is the hypothesis variable, values $h_1, h_2, \dots$, prior $\mathbf{P}(H)$
 j th observation d_j gives the outcome of random variable D_j
training data $\mathbf{d} = d_1, \dots, d_N$

Given the data so far, each hypothesis has a posterior probability:

$$P(h_i|\mathbf{d}) = \alpha P(\mathbf{d}|h_i)P(h_i)$$

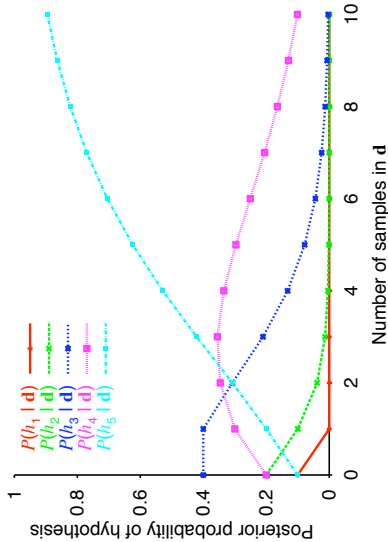
where $P(\mathbf{d}|h_i)$ is called the **likelihood**

Predictions use a likelihood-weighted average over the hypotheses:

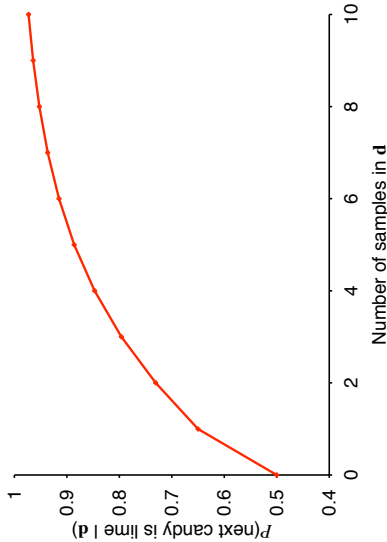
$$\mathbf{P}(X|\mathbf{d}) = \sum_i \mathbf{P}(X|\mathbf{d}, h_i)P(h_i|\mathbf{d}) = \sum_i \mathbf{P}(X|h_i)P(h_i|\mathbf{d})$$

No need to pick one best-guess hypothesis!

Posterior probability of hypotheses



Prediction probability



ML approximation

For large data sets, prior becomes irrelevant

Maximum likelihood (ML) learning: choose h_{ML} maximizing $P(\mathbf{d}|h_u)$

i.e., simply get the best fit to the data; identical to MAP for uniform prior (which is reasonable if all hypotheses are of the same complexity)

ML is the “standard” (non-Bayesian) statistical learning method

Multiple parameters contd.

Derivatives of L contain only the relevant parameter:

$$\frac{\partial L}{\partial \theta} = \frac{c}{\theta} - \frac{\ell}{1 - \theta} = 0 \quad \Rightarrow \quad \theta = \frac{c}{c + \ell}$$

$$\frac{\partial L}{\partial \theta_1} = \frac{r_c}{\theta_1} - \frac{g_c}{1 - \theta_1} = 0 \quad \Rightarrow \quad \theta_1 = \frac{r_c}{r_c + g_c}$$

$$\frac{\partial L}{\partial \theta_2} = \frac{r_\ell}{\theta_2} - \frac{g_\ell}{1 - \theta_2} = 0 \quad \Rightarrow \quad \theta_2 = \frac{r_\ell}{r_\ell + g_\ell}$$

With complete data, parameters can be learned separately

ML parameter learning in Bayes nets

Bag from a new manufacturer; fraction θ of cherry candies?

Any θ is possible: continuum of hypotheses h_θ

θ is a parameter for this simple (binomial) family of models

Suppose we unwrap N candies, c cherries and $\ell = N - c$ limes

These are i.i.d. (independent, identically distributed) observations, so

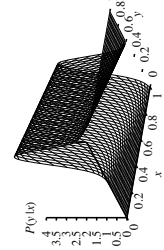
$$P(\mathbf{d}|h_\theta) = \prod_{j=1}^N P(d_j|h_\theta) = \theta^c \cdot (1 - \theta)^\ell$$

Maximize this w.r.t. θ —which is easier for the log-likelihood:

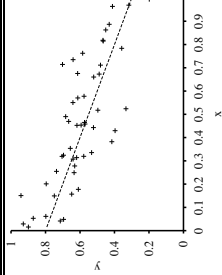
$$L(\mathbf{d}|h_\theta) = \log P(\mathbf{d}|h_\theta) = \sum_{j=1}^N \log P(d_j|h_\theta) = c \log \theta + \ell \log(1 - \theta)$$

$$\frac{dL(\mathbf{d}|h_\theta)}{d\theta} = \frac{c}{\theta} - \frac{\ell}{1 - \theta} = 0 \quad \Rightarrow \quad \theta = \frac{c}{c + \ell} = \frac{c}{N}$$

Seems sensible, but causes problems with 0 counts!



Example: linear Gaussian model



$$\text{Maximizing } P(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y - (\theta_1 x + \theta_2))^2}{2\sigma^2}} \text{ w.r.t. } \theta_1, \theta_2$$

$$= \text{minimizing } E = \sum_{j=1}^N (y_j - (\theta_1 x_j + \theta_2))^2$$

That is, minimizing the sum of squared errors gives the ML solution for a linear fit assuming Gaussian noise of fixed variance

