430.457

Introduction to Intelligent Systems

Prof. Songhwai Oh

ECE, SNU

Bayesian Learning

Bayesian Learning

- Calculates the probability of each hypothesis, given the data.
- Makes predictions based on all the hypotheses, weighted by their probabilities.
- Probability of hypothesis h_i given data $\mathbf{d}$ (Bayes' rule):

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

- $P(h_i)$: **hypothesis prior**
- $P(\mathbf{d}|h_i)$: **likelihood** of the data under hypothesis h_i.
- The likelihood of the data is calculated under the assumption that the observations are i.i.d.

$$P(\mathbf{d}|h_i) = \prod_j P(d_j|h_i).$$

- Prediction about an unknown quantity X:

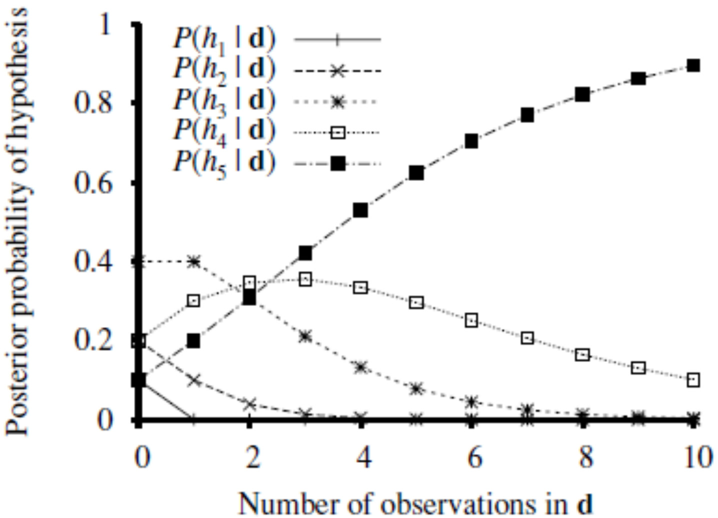
$$P(X|\mathbf{d}) = \sum_i P(X|\mathbf{d}, h_i)P(h_i|\mathbf{d}) = \sum_i P(X|h_i)P(h_i|\mathbf{d}).$$

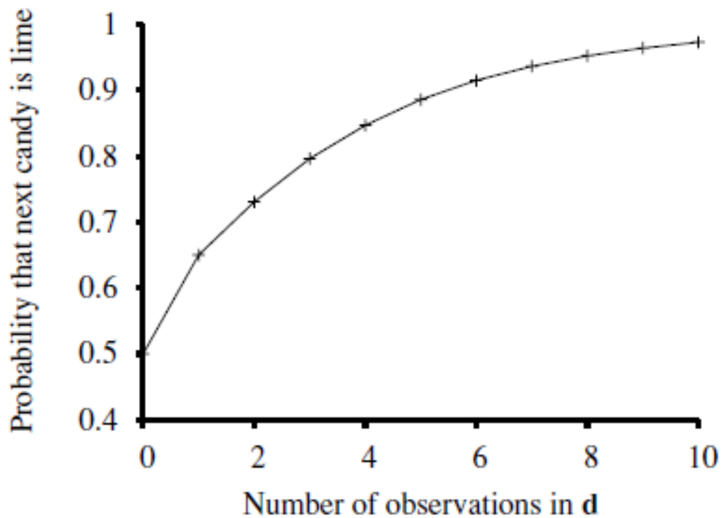
Example

Our favorite Surprise candy comes in two flavors: cherry (yum) and lime (ugh). The manufacturer has a peculiar sense of humor and wraps each piece of candy in the same opaque wrapper, regardless of flavor. The candy is sold in very large bags, of which there are known to be five kinds – indistinguishable from the outside:

h_1: 100% cherry,
h_2: 75% cherry + 25% lime,
h_3: 50% cherry + 50% lime,
h_4: 25% cherry + 75% lime,
h_5: 100% lime .

Suppose the bag is really an all-lime bag (h_5) and the first 10 candies are all lime; then $P(\mathbf{d}|h_3) = 0.5^{10}$





MAP and ML

- The Bayesian prediction is *optimal*.
- But finding a solution to the Bayesian prediction problem is computationally expensive when the hypothesis space is large or infinite (cannot do the summation or integration).
- We usually use approximate or simplified solutions.
- **Maximum a posteriori (MAP)**: Prediction based on the MAP hypothesis h_{MAP}, where

$$h_{MAP} = \arg\max_{h \in \mathcal{H}} P(h|\mathbf{d}) = \arg\max_{h \in \mathcal{H}} P(\mathbf{d}|h)P(h).$$

- **Maximum likelihood (ML)**: Assume a uniform prior over $\mathcal{H}$. Then the MAP learning is reduced to finding h that maximizes the likelihood.

$$h_{ML} = \arg\max_{h \in \mathcal{H}} P(\mathbf{d}|h)P(h) = \arg\max_{h \in \mathcal{H}} P(\mathbf{d}|h).$$