CSC421 Intro to Artificial Intelligence

UNIT 31:
Statistical Learning

Outline

- Baysian Learning
- Maximum a posteriori (MAP) and Maximum Likelihood Learning
- ML parameter learning with complete data

Full Baysian Learning

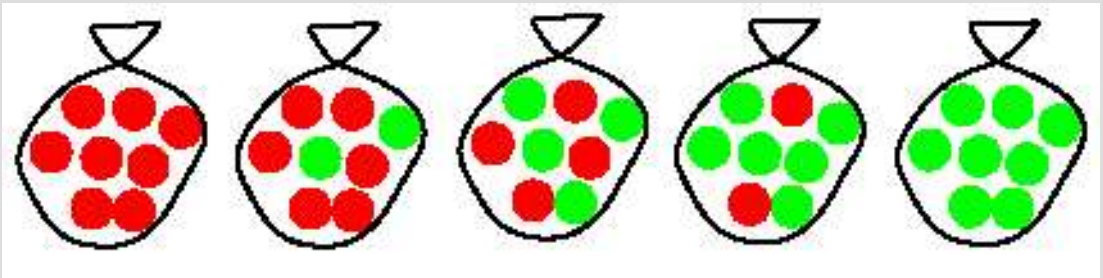
- View learning as baysian updating of a probability distribution over the hypothesis space
- H is the hypothesis variable, $h_1,..,h_n$, prior P(H)
- jth observation gives outcome of rv D_j
  - Training data $\underline{d} = d_1,...,d_n$
- Given the data so far, each hypothesis has a posterior probability
  - $P(h_i \mid \underline{d}) = ?$

Full Baysian Learning

- $P(h_i \mid \underline{d}) = a\, P(\underline{d} \mid h_i)\, P(h_i)$ where $P(\underline{d} \mid h_i)$ is called the likelihood
- Predictions use a likelihood weighted average over the hypotheses:
  - $P(X \mid \underline{d}) = \sum_i P(X \mid \underline{d}, h_i)\, P(h_i \mid \underline{d}) = \sum_i P(X \mid h_i)\, P(h_i \mid \underline{d})$
  - Assume each hypothesis determines a probability distribution over X
  - No need to pick best guess
- IID (independently and identically distributed):
  - $P(\underline{d} \mid hi) = \prod_j P(d_j \mid h_i)$

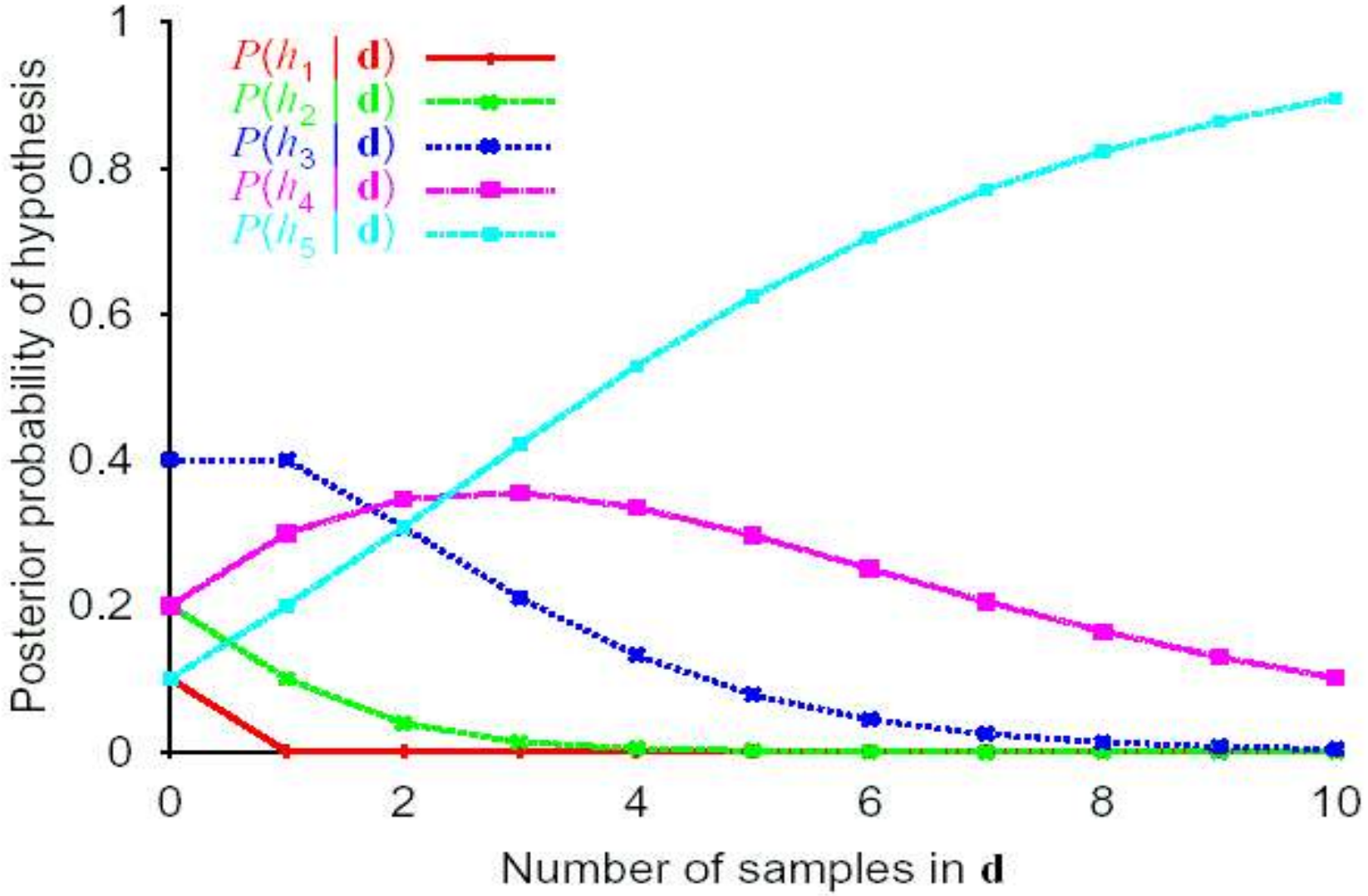
Example

- Suppose there are 5 kinds of of bags of candies :
  - 10% are h_1: 100% cherry candies
  - 20% are h2: 75% cherry candies 25% lime
  - 40% are h3: 50% cherry 50% lime
  - 20% are h4: 25% cherry 75% lime
  - 10% are h5: 50% cherry 50% lime
- Observe



What kind of bag is it ?
What flavour will be the next candy ?

Posterior Probabilities of Hypotheses

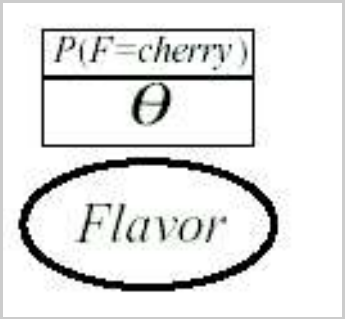


MAP approximation

- Summing over entire hypothesis space intractable
- Maximum a Posteriori (MAP) learning: choose h_{MAP} maximizing $P(h_i | \underline{d})$
  - i.e. Maximize $P(\underline{d}|h_i)\ P(h_i)$ or $\log P(d|h_i) + \log P(h_i)$
  - Log terms can be viewed as (negative of)
    - Bits to encode data/hypothesis + bits to encode hypothesis

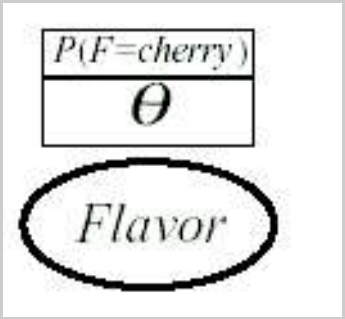
ML approximation

- For large datasets prior becomes irrelevant
- Maximum likelihood (ML) learning: choose h_{ML} maximizing $P(\underline{d}|h_i)$
  - i.e. Simply get the best fit for the data – identical to MAP with uniform prior
- ML is the "standard" (non-Baysian) statistical learning method



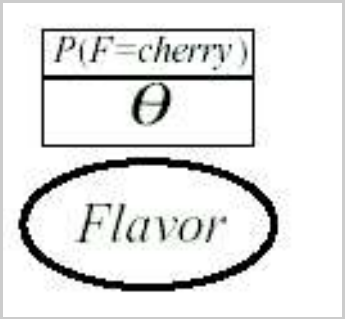
ML parameter learning in Bayes nets

- Bag from a new manufactorer; fraction θ of cherry candies ? Any θ is possible – continuum of hypotheses – θ is a parameter for this simple (binomial) family of models
- Suppose we unwrap N candies, c cherries and l = N – c limes. These are IID:
  - $P(\underline{d}|\, h_\theta) = ?$



ML parameter learning in Bayes nets

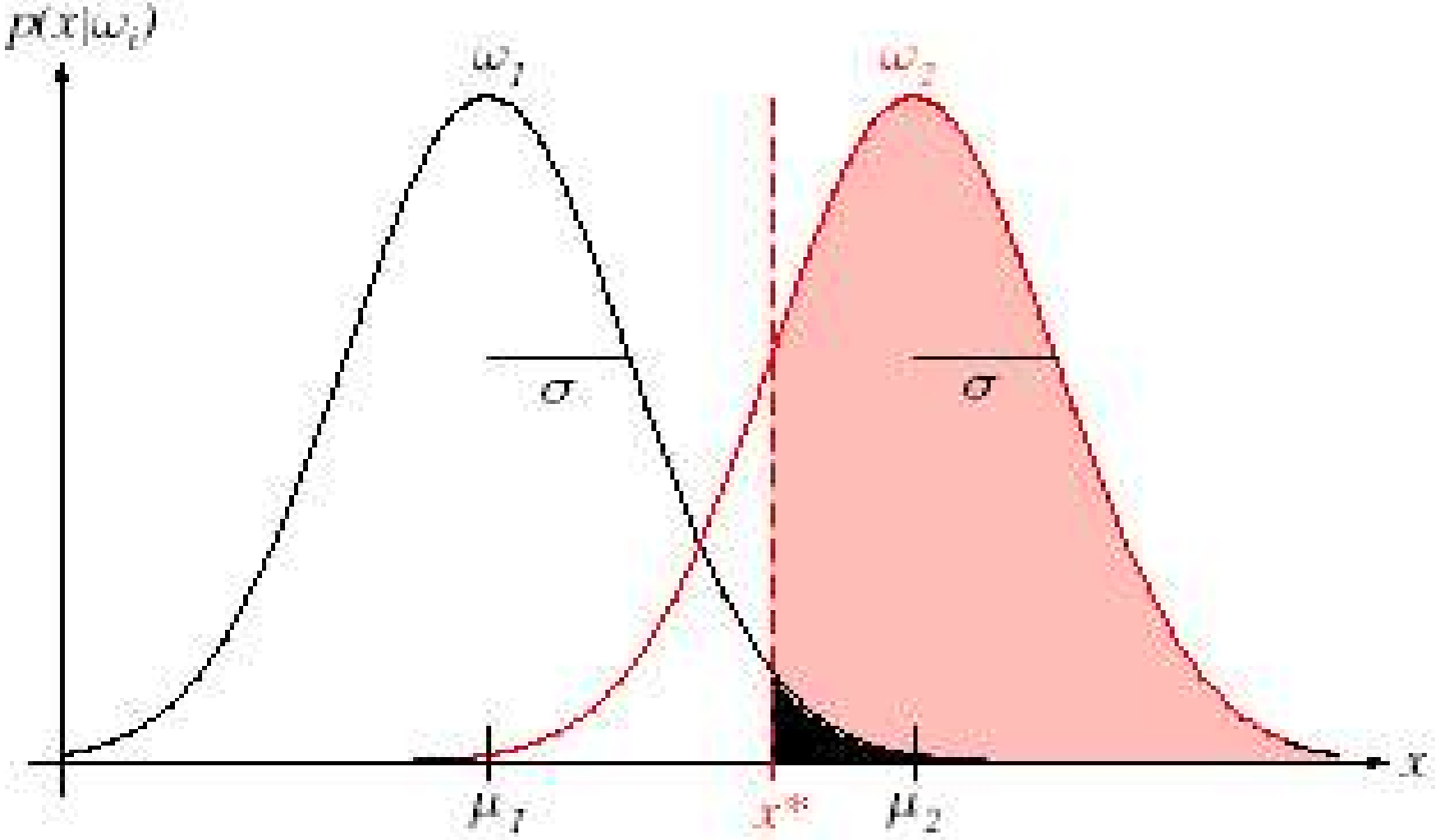
- Suppose we unwrap N candies, c cherries and l = N – c limes. These are IID:
  - $P(\underline{d} \mid h_\theta) = \prod P(dj \mid h_\theta) = \theta^c (1-\theta)^c$
- Maximize with respect to θ log likelihood
  - $L(\underline{d} \mid \theta) = \log P(\underline{d} \mid h_\theta) = \sum \log P(d_j \mid h_\theta) = c \log \theta + l \log(1-\theta)$
  - $dL(\underline{d} \mid \theta) / d\theta) = ?$



ML parameter learning in Bayes nets

- Suppose we unwrap N candies, c cherries and l = N – c limes. These are IID:
  - $P(\underline{d} \mid h_\theta) = \prod P(dj \mid h_\theta) = \theta^c (1-\theta)^c$
- Maximize with respect to θ log likelihood
  - $L(\underline{d} \mid \theta) = \log P(\underline{d} \mid h_\theta) = \sum \log P(d_j \mid h_\theta) = c \log \theta + l \log(1-\theta)$
  - $dL(\underline{d} \mid \theta) / d\theta) = c/\theta - l/1-\theta = 0$
  - $\Rightarrow \theta = c / c + l = c / N$

Classification using single Gaussian



ML Summary

- ML steps
  - Choose parametrized family of functions
  - Write down likelihood of data as a function of parameters
  - Write derivative of log likelihood
  - Find parameters θ such that derivative is zero
    - Analytically or numerically (optimization)