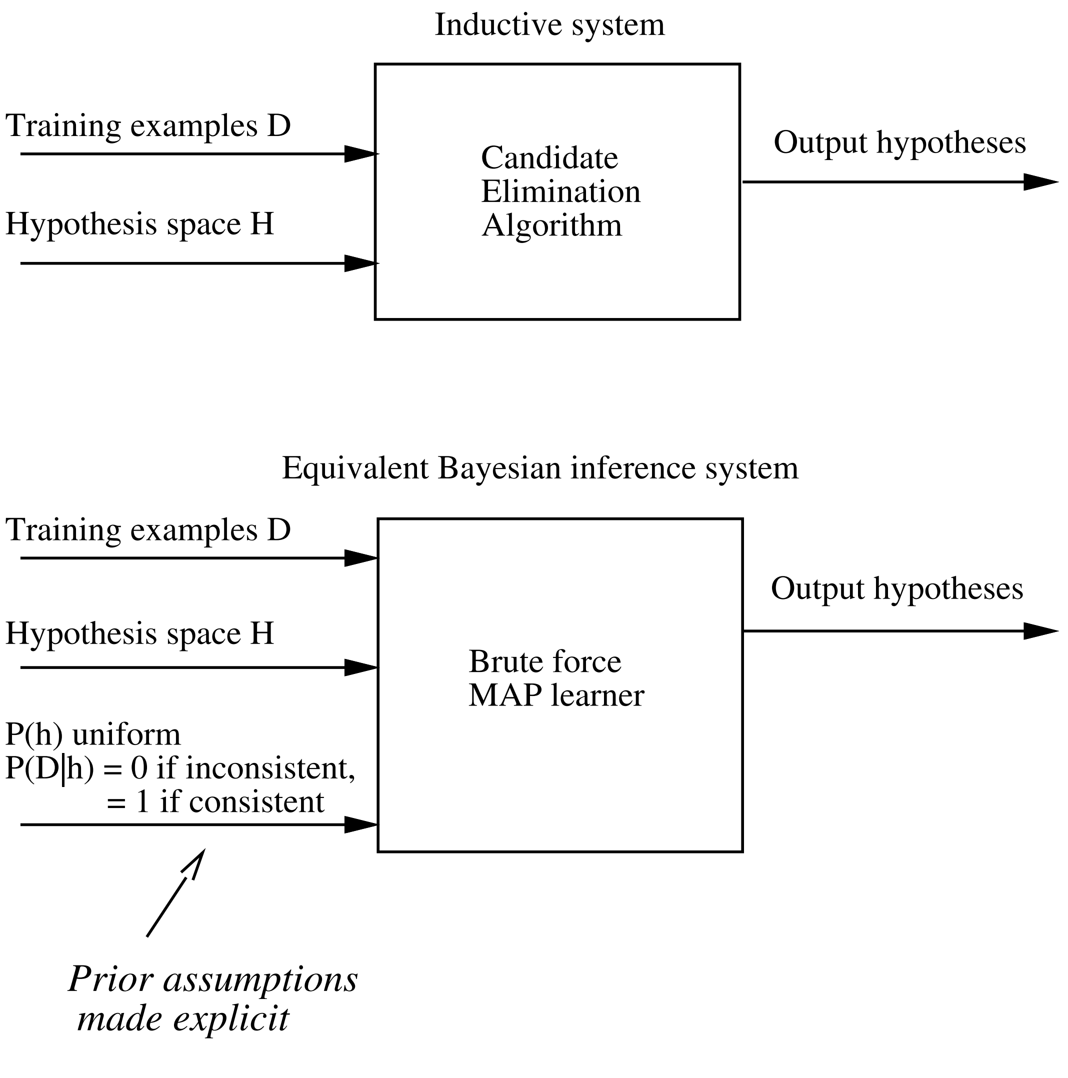
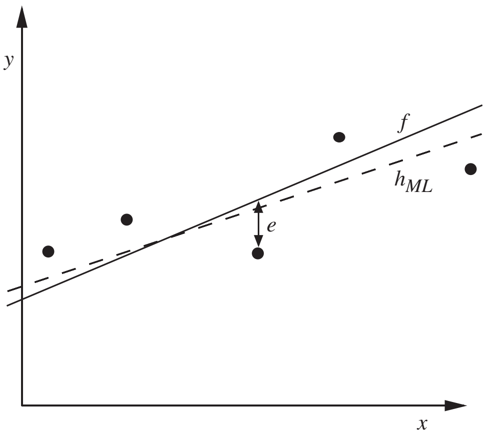
# Characterizing Learning Algorithms by Equivalent MAP Learners



# Learning A Real Valued Function




Consider any real-valued target function $f$

Training examples $\langle x_i, d_i \rangle$, where $d_i$ is noisy training value

- $d_i = f(x_i) + e_i$
- $e_i$ is random variable (noise) drawn independently for each $x_i$ according to some Gaussian distribution with mean=0

Then the maximum likelihood hypothesis $h_{ML}$ is the one that minimizes the sum of squared errors:

$$h_{ML} = \arg\min_{h \in H} \sum_{i=1}^{m} (d_i - h(x_i))^2$$

# Learning A Real Valued Function

---

$$\begin{aligned}
h_{ML} &= \underset{h \in H}{\operatorname{argmax}}\, p(D|h) \\
&= \underset{h \in H}{\operatorname{argmax}} \prod_{i=1}^{m} p(d_i|h) \\
&= \underset{h \in H}{\operatorname{argmax}} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{d_i - h(x_i)}{\sigma})^2}
\end{aligned}$$

Maximize natural log of this instead...

$$\begin{aligned}
h_{ML} &= \underset{h \in H}{\operatorname{argmax}} \sum_{i=1}^{m} \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2 \\
&= \underset{h \in H}{\operatorname{argmax}} \sum_{i=1}^{m} -\frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2 \\
&= \underset{h \in H}{\operatorname{argmax}} \sum_{i=1}^{m} -(d_i - h(x_i))^2 \\
&= \underset{h \in H}{\operatorname{argmin}} \sum_{i=1}^{m} (d_i - h(x_i))^2
\end{aligned}$$

# Learning to Predict Probabilities

Consider predicting survival probability from patient data

Training examples $\langle x_i, d_i \rangle$, where $d_i$ is 1 or 0

Want to train neural network to output a *probability* given $x_i$ (not a 0 or 1)

In this case can show

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}} \sum_{i=1}^{m} d_i \ln h(x_i) + (1 - d_i) \ln(1 - h(x_i))$$

Weight update rule for a sigmoid unit:

$$w_{jk} \leftarrow w_{jk} + \Delta w_{jk}$$

where

$$\Delta w_{jk} = \eta \sum_{i=1}^{m} (d_i - h(x_i)) \; x_{ijk}$$

# Minimum Description Length Principle

Occam's razor: prefer the shortest hypothesis

MDL: prefer the hypothesis $h$ that minimizes

$$h_{MDL} = \operatorname*{argmin}_{h \in H} L_{C_1}(h) + L_{C_2}(D|h)$$

where $L_C(x)$ is the description length of $x$ under encoding $C$

Example: $H$ = decision trees, $D$ = training data labels

- $L_{C_1}(h)$ is # bits to describe tree $h$
- $L_{C_2}(D|h)$ is # bits to describe $D$ given $h$
  - Note $L_{C_2}(D|h) = 0$ if examples classified perfectly by $h$. Need only describe exceptions
- Hence $h_{MDL}$ trades off tree size for training errors

# Minimum Description Length Principle

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} P(D|h)P(h) \\ &= \arg\max_{h \in H} \log_2 P(D|h) + \log_2 P(h) \\ &= \arg\min_{h \in H} -\log_2 P(D|h) - \log_2 P(h) \quad (1) \end{aligned}$$

Interesting fact from information theory:

The optimal (shortest expected coding length) code for an event with probability $p$ is $-\log_2 p$ bits.

So interpret (1):

- $-\log_2 P(h)$ is length of $h$ under optimal code
- $-\log_2 P(D|h)$ is length of $D$ given $h$ under optimal code (if assume $x_i$ fixed or indep of $h$, this is just length of encoding the misclassifications)

$\rightarrow$ prefer the hypothesis that minimizes

$$length(h) + length(misclassifications)$$

## Most Probable Classification of New Instances

---

So far we've sought the most probable *hypothesis* given the data $D$ (i.e., $h_{MAP}$)

Given new instance $x$, what is its most probable *classification*?

- $h_{MAP}(x)$ is not the most probable classification!

Consider:

- Three possible hypotheses:

$$P(h_1|D) = .4,\ P(h_2|D) = .3,\ P(h_3|D) = .3$$

- Given new instance $x$,

$$h_1(x) = +,\ h_2(x) = -,\ h_3(x) = -$$

- What's most probable classification of $x$?

# Bayes Optimal Classifier

**Bayes optimal classification:**

$$\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

Example:

$$\begin{array}{l} P(h_1|D) = .4,\ P(-|h_1) = 0,\ P(+|h_1) = 1 \\ P(h_2|D) = .3,\ P(-|h_2) = 1,\ P(+|h_2) = 0 \\ P(h_3|D) = .3,\ P(-|h_3) = 1,\ P(+|h_3) = 0 \end{array}$$

therefore

$$\begin{aligned} \sum_{h_i \in H} P(+|h_i)P(h_i|D) &= .4 \\ \sum_{h_i \in H} P(-|h_i)P(h_i|D) &= .6 \end{aligned}$$

and

$$\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D) = -$$

# Gibbs Classifier

Bayes optimal classifier provides best result, but can be expensive if many hypotheses.
Gibbs algorithm:

1. Choose one hypothesis at random, according to $P(h|D)$
2. Use this to classify new instance

Surprising fact: Assume target concepts are drawn at random from $H$ according to priors on $H$. Then:

$$E[error_{Gibbs}] \leq 2E[error_{BayesOptimal}]$$

Suppose correct, uniform prior distribution over $H$, then

- Pick any hypothesis from VS, with uniform probability
- Its expected error no worse than twice Bayes optimal