Learning Objectives

At the end of the class you should be able to:

- derive Bayesian learning from first principles
- explain how the Beta and Dirichlet distributions are used for Bayesian learning.

Model Averaging (Bayesian Learning)

We want to predict the output Y of a new case that has input $X = x$ given the training examples $\mathbf{e}$:

$$
\begin{aligned}
p(Y|x \wedge \mathbf{e}) &= \sum_{m \in M} P(Y \wedge m | x \wedge \mathbf{e}) \\
&=
\end{aligned}
$$

Model Averaging (Bayesian Learning)

We want to predict the output Y of a new case that has input $X = x$ given the training examples $\mathbf{e}$:

$$
\begin{aligned}
p(Y|x \wedge \mathbf{e}) &= \sum_{m \in M} P(Y \wedge m|x \wedge \mathbf{e}) \\
&= \sum_{m \in M} P(Y|m \wedge x \wedge \mathbf{e}) P(m|x \wedge \mathbf{e}) \\
&=
\end{aligned}
$$

Model Averaging (Bayesian Learning)

We want to predict the output Y of a new case that has input $X = x$ given the training examples $\mathbf{e}$:

$$
\begin{aligned}
p(Y|x \wedge \mathbf{e}) &= \sum_{m \in M} P(Y \wedge m | x \wedge \mathbf{e}) \\
&= \sum_{m \in M} P(Y|m \wedge x \wedge \mathbf{e}) P(m|x \wedge \mathbf{e}) \\
&= \sum_{m \in M} P(Y|m \wedge x) P(m|\mathbf{e})
\end{aligned}
$$

M is a set of mutually exclusive and covering hypotheses.

- What assumptions are made here?

Learning Under Uncertainty

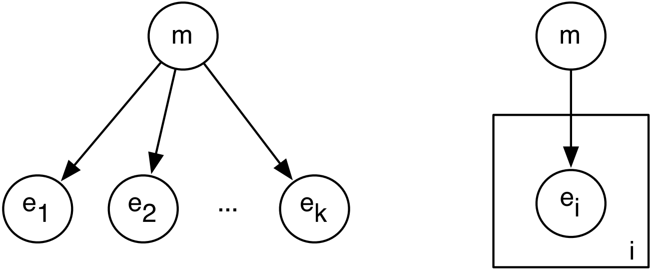
- The posterior probability of a model given examples **e**:

$$P(m|\mathbf{e}) = \frac{P(\mathbf{e}|m) \times P(m)}{P(\mathbf{e})}$$

- The likelihood, $P(\mathbf{e}|m)$, is the probability that model m would have produced examples **e**.
- The prior, $P(m)$, encodes the learning bias
- $P(\mathbf{e})$ is a normalizing constant so the probabilities of the models sum to 1.
- Examples $\mathbf{e} = \{e_1, \ldots, e_k\}$ are independent and identically distributed (i.i.d.) given m if

$$P(\mathbf{e}|m) = \prod_{i=1}^{k} P(e_i|m)$$

Plate Notation



Bayesian Leaning of Probabilities

- Y has two outcomes y and $\neg y$.
  We want the probability of y given training examples $\mathbf{e}$.
- We can treat the probability of y as a real-valued random variable on the interval $[0, 1]$, called ϕ. Bayes' rule gives:

$$P(\phi{=}p|\mathbf{e}) =$$

Bayesian Leaning of Probabilities

- Y has two outcomes y and $\neg y$.
  We want the probability of y given training examples **e**.
- We can treat the probability of y as a real-valued random variable on the interval $[0, 1]$, called ϕ. Bayes' rule gives:

$$P(\phi{=}p|\mathbf{e}) = \frac{P(\mathbf{e}|\phi{=}p) \times P(\phi{=}p)}{P(\mathbf{e})}$$

- Suppose **e** is a sequence of n_1 instances of y and n_0 instances of $\neg y$:

$$P(\mathbf{e}|\phi{=}p) =$$

Bayesian Leaning of Probabilities

- Y has two outcomes y and $\neg y$.
  We want the probability of y given training examples $\mathbf{e}$.
- We can treat the probability of y as a real-valued random variable on the interval $[0,1]$, called ϕ. Bayes' rule gives:

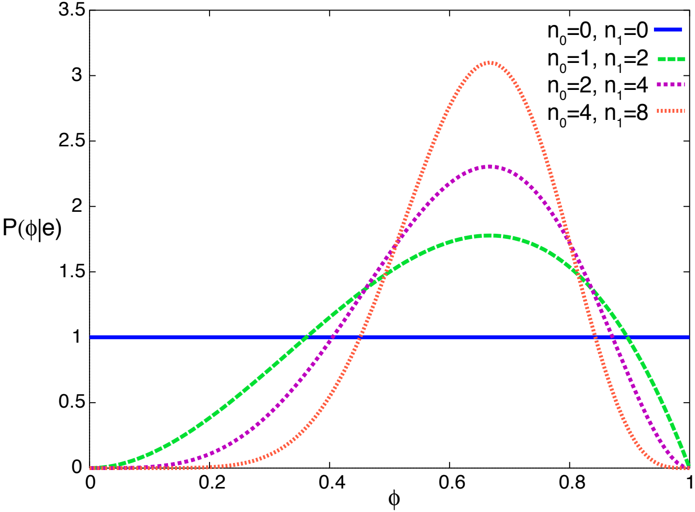
$$P(\phi{=}p|\mathbf{e}) = \frac{P(\mathbf{e}|\phi{=}p) \times P(\phi{=}p)}{P(\mathbf{e})}$$

- Suppose $\mathbf{e}$ is a sequence of n_1 instances of y and n_0 instances of $\neg y$:

$$P(\mathbf{e}|\phi{=}p) = p^{n_1} \times (1-p)^{n_0}$$

- Uniform prior: $P(\phi{=}p) = 1$ for all $p \in [0,1]$.

Posterior Probabilities for Different Training Examples (beta distribution)



MAP model

- The maximum a posteriori probability (MAP) model is the model m that maximizes $P(m|\mathbf{e})$. That is, it maximizes:

$$P(\mathbf{e}|m) \times P(m)$$

- Thus it minimizes:

$$(-\log P(\mathbf{e}|m)) + (-\log P(m))$$

which is the number of bits to send the examples, $\mathbf{e}$, given the model m plus the number of bits to send the model m.

Averaging Over Models

- Idea: Rather than choosing the most likely model, average over all models, weighted by their posterior probabilities given the examples.
- If you have observed a sequence of n_1 instances of y and n_0 instances of $\neg y$, with uniform prior:
  - the most likely value (MAP) is $\dfrac{n_1}{n_0 + n_1}$
  - the expected value is $\dfrac{n_1 + 1}{n_0 + n_1 + 2}$

Beta Distribution

$$Beta^{\alpha_0,\alpha_1}(p) = \frac{1}{K} p^{\alpha_1 - 1} \times (1-p)^{\alpha_0 - 1}$$

where K is a normalizing constant. $\alpha_i > 0$.

- The uniform distribution on $[0,1]$ is $Beta^{1,1}$.
- The expected value is $\alpha_1/(\alpha_0 + \alpha_1)$.

If the prior probability of a Boolean variable is $Beta^{\alpha_0,\alpha_1}$, the posterior distribution after observing n_1 true cases and n_0 false cases is:

Beta Distribution

$$Beta^{\alpha_0,\alpha_1}(p) = \frac{1}{K} p^{\alpha_1 - 1} \times (1-p)^{\alpha_0 - 1}$$

where K is a normalizing constant. $\alpha_i > 0$.

- The uniform distribution on $[0,1]$ is $Beta^{1,1}$.
- The expected value is $\alpha_1/(\alpha_0 + \alpha_1)$.

If the prior probability of a Boolean variable is $Beta^{\alpha_0,\alpha_1}$, the posterior distribution after observing n_1 true cases and n_0 false cases is:

$$Beta^{\alpha_0+n_0,\alpha_1+n_1}$$

Dirichlet distribution

- Suppose Y has k values.
- The Dirichlet distribution has two sorts of parameters,
  - positive counts $\alpha_1, \ldots, \alpha_k$
    α_i is one more than the count of the ith outcome.
  - probability parameters $p_1, \ldots, p_k$
    p_i is the probability of the ith outcome

$$Dirichlet^{\alpha_1,\ldots,\alpha_k}(p_1, \ldots, p_k) = \frac{1}{K} \prod_{j=1}^{k} p_j^{\alpha_j - 1}$$

  where K is a normalizing constant
- The expected value of ith outcome is

$$\frac{\alpha_i}{\sum_j \alpha_j}$$

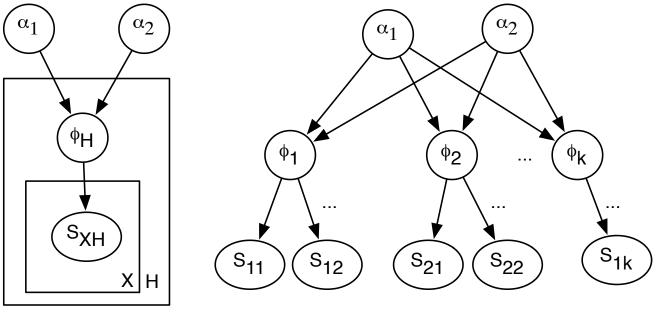
Hierarchical Bayesian Model

Where do the priors come from?

Example: S_{XH} is true when patient X is sick in hospital H.

We want to learn the probability of Sick for each hospital.

Where do the prior probabilities for the hospitals come from?



(a) (b)