

Model Averaging

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

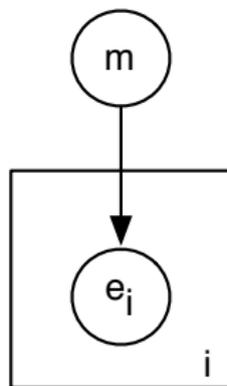
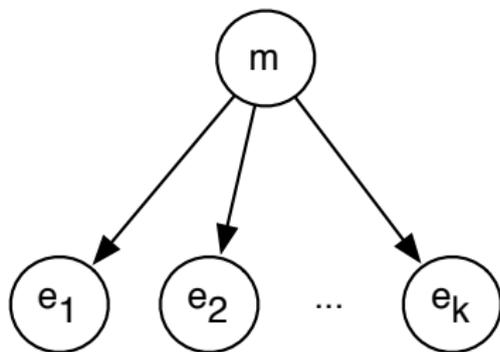
- The posterior probability of a model given examples \mathbf{e} :

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

- The **likelihood**, $P(\mathbf{e}|m)$, is the probability that model m would have produced examples \mathbf{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 $\mathbf{e} = \{e_1, \dots, e_k\}$ 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 Learning 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 .

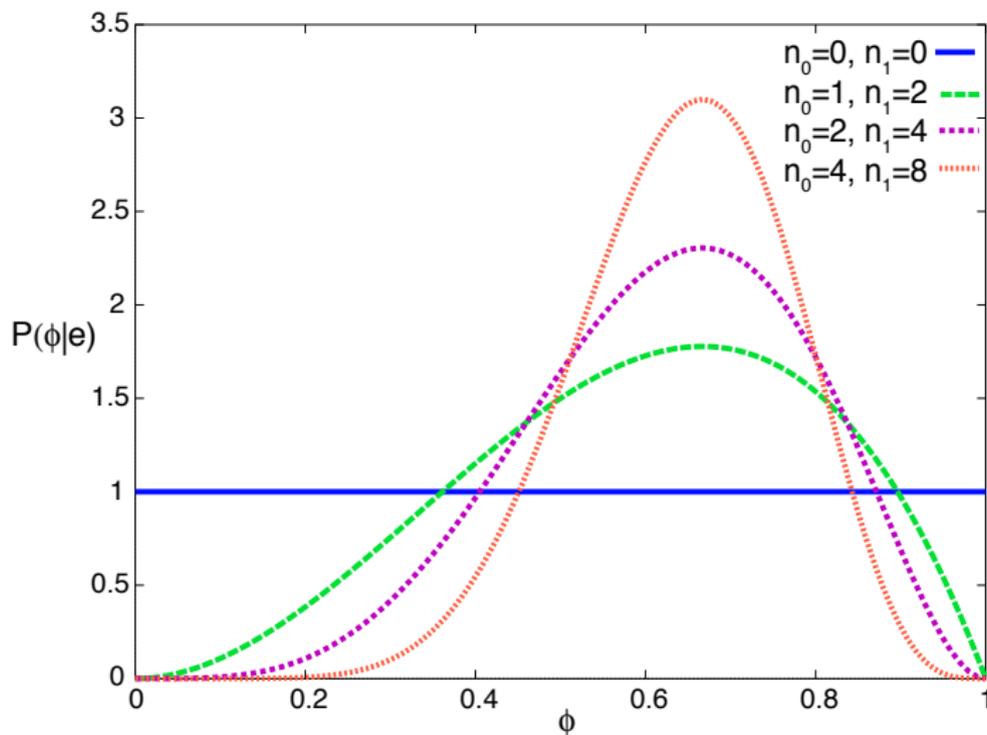
$$P(\phi=p|\mathbf{e}) = \frac{P(\mathbf{e}|\phi=p) \times P(\phi=p)}{P(\mathbf{e})}$$

- 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)



- 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 $\frac{n_1}{n_0+n_1}$
 - ▶ the expected value is $\frac{n_1+1}{n_0+n_1+2}$

Beta Distribution

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

where K is a normalizing constant.

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

Dirichlet distribution

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

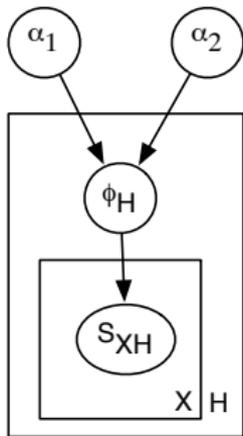
$$\text{Dirichlet}^{\alpha_1, \dots, \alpha_k}(p_1, \dots, p_k) = \frac{1}{K} \prod_{j=1}^k p_j^{\alpha_j - 1}$$

where K is a normalizing constant

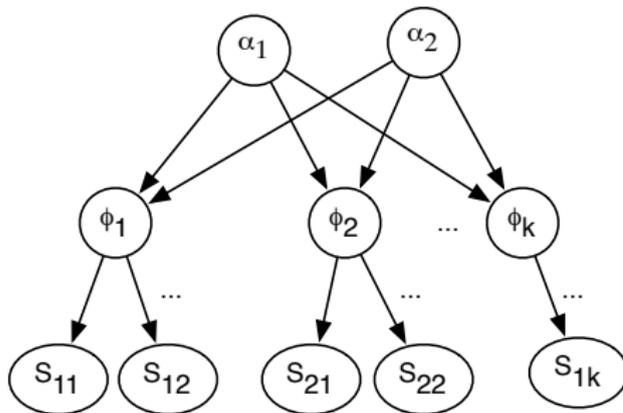
- The expected value of i th outcome is

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

Hierarchical Bayesian Model



(a)



(b)