Bayes' Formula

Query problem (lecture)

$$P(R_d|Q) = \frac{P(Q|R_d)P(R_d)}{P(R_d)}$$

- $R_d = 1$ if document d relevant, $Q = (q_1, \ldots, q_n)$ query.
- Recall: Unigram assumption, so $P(Q|R_d) = \prod_{i=1}^{n} P(q_i|R_d)$.

Bayes' formula: Schematic

$$P(\text{State}|\text{Observation}) = \frac{P(\text{Observation}|\text{State})P(\text{State})}{P(\text{Observation})}$$

State = Random value to estimate from dependence on observation.

Components

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

Meaning of Components

Components

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

- **Posterior:** The model to compute.
- **Evidence:** Given Q, evidence is constant, not important for comparison of different d (same Q).
  - Neglect evidence: For comparison between different d for same Q (e.g. retrieve d_1 or d_2?). Suitable for ranking.
  - Required: For absolute scoring (e.g. retrieve all documents with $P(R_d|Q) >$ threshold)
- **Likelihood:** $P(Q|R_d)$ means: How likely would it be to generate the query Q we have observed, if we knew that d is relevant?
- **Prior:** Probability of the document being relevant "in general". May be based on any information about the documents available when model is computed, but *not on the query*.

Likelihood Idea

Problem setting

- Given: Two "states" y_1, y_2, observation x, likelihood $P(x|Y)$.
- Examples of states: Relevance above, model/parameter etc
- Given that x has been observed, which one should we choose?

Maximum likelihood principle

Choose the one which makes x more probable to happen.

Estimation strategies

1. Pick the state that maximizes the likelihood.
2. Pick the state that maximizes the posterior.

Likelihood only vs. full Bayesian formula

Likelihood models generation of observation given state; does not take into account information about state.