Bayesian Learning

Arnav Gupta

November 27, 2024

Contents

1 Bayesian Learning

Premise:

- have a number of hypotheses or models
- assume all are correct to some degree
- have a distribution over the models
- compute expected prediction given this average

For input features X, target feature Y, and evidence $d = \{x_1, y_1, \ldots, x_N, y_N\}$, with new input x to find the corresponding output y sum over all models

$$P(Y \mid x, d) = \sum_{m \in M} P(Y \mid m, x) P(m \mid d)$$

For a hypothesis H, the **prior** is $P(H)$ and the **likelihood** is $P(d \mid H)$. With **Bayesian Learning**, update the posterior (Bayes theorem)

$$P(H \mid d) \propto P(d \mid H)P(H)$$

To predict X

$$P(X \mid d) = \sum_i P(X \mid h_i)P(h_i \mid d)$$

Predictions are weighted averages of the predictions of the individual hypotheses. The hypotheses serve as intermediaries between raw data and predictions.

Properties of Bayesian learning:

- **optional**: given the prior, no other prediction is correct more often than the Bayesian one
- **no overfitting**: the prior and likelihood both penalize complex hypotheses

Bayesian learning may be unmanageable when the hypothesis space is large. Further, the sum over hypotheses space may be unmanageable. The solution is to approximate Bayesian learning with other strategies.

2 Maximum a Posteriori (MAP)

Make predictions based on the most probable hypothesis h_{MAP} where

$$h_{MAP} = \text{argmax}_{h_i} P(h_i \mid d)$$

Note that

$$P(X \mid d) \approx P(X \mid h_{MAP})$$

Less accurate than full Bayesian, but converges as data increases. No overfitting like Bayesian. Finding h_{MAP} may be unmanageable since the following induces nonlinear optimization

$$h_{MAP} = \text{argmax}_h P(h) \prod_i P(d_i \mid h)$$

Taking the log can linearize this.

3 Maximum Likelihood (ML)

Simplify MAP by assuming uniform prior:

$$h_{ML} = \text{argmax}_h P(d \mid h)$$

so make prediction based solely on h_{ML}:

$$P(X \mid d) \approx P(X \mid h_{ML})$$

Less accurate than full Bayesian or MAP, but converges as data increases. More susceptible to overfitting since no prior. h_{ML} is easier to find than h_{MAP}, especially taking the log.

3.1 Binomial Distribution

Generalize the hypothesis space to a continuous quantity using the binomial distribution.

For priors on binomials, the **Beta distribution** is

$$B(\theta; a, b) \propto \theta^{a-1}(1-\theta)^{b-1}$$

4 Classifiers

If the classification is known, the feature values could be predicted

$$P(\text{Class} \mid X_1 \dots X_n) \propto P(X_1, \dots, X_n \mid \text{Class})P(\text{Class})$$

With a **Naive Bayesian classifier** X_i are independent of each other given the class, though this requires $P(\text{Class})$ and $P(X_i \mid \text{Class})$ for each X_i.

4.1 Laplace Correction

If a feature never occurs in the training set, but it does in the test set, ML assigns 0 probability to a high likelihood class.

To correct for this, add 1 to the numerator and d to the denominator, where d is the arity of the variable.

4.2 Bayesian Network Parameter Learning

For fully observed data with:

- parameters $\theta_{V,pa(V)=v^i}$
- CPTs $\theta_{V,pa(V)=v} = P(V \mid pa(V) = v)$
- data d where

$$d_i = \langle V_1 = v_{1,i}, \ldots, V_n = v_{n,i} \rangle$$

To get the maximum likelihood, set $\theta_{V,pa(V)=v}$ to the relative frequency of values of V given the values v of the parents of V.

4.3 Occam's Razor

Simplicity is encouraged in the likelihood function. A hypothesis that is more complex (lower bias) can explain more datasets but it can only explain each with lower probability (higher variance).