Bayesian Learning

1 Hypothesis analysis

Using Bayes rule, we can reason about the "best" hypothesis in a hypothesis class given some data. From a probabilistic standpoint, "best" is defined as most likely. With probabilities of hypotheses h in a hypothesis class $\mathcal{H}$ given some data S, we can then define the best hypothesis h^* as

$$h^* = \arg\max_h p(h \mid S)$$

Applying Bayes rule, we can rewrite this in a more usable form, as such

$$p(h \mid S) = \frac{p(S \mid h)p(h)}{p(S)}$$

The terms in the right-hand side are much more manageable and meaningful, allowing us to reason about the left-hand side.

$p(h)$ and $p(S)$ are the priors: $p(h)$ describes the probability of a particular hypothesis and encodes our inductive bias (for example, that simpler polynomials are more likely than complex polynomials or that smaller decision trees are more likely than bigger decision trees); $p(S)$ describes the likelihood of the data itself, but is usually assumed to be uniform and is therefore ignored for the most part since it is a constant term (it doesn't depend on h) and won't be useful for further analyses.

$p(S \mid h)$ is the posterior and describes the likelihood that the data S would be generated by some underlying process which includes h.

Since we want to find h^* by maximizing $p(h \mid S)$ which equals $\frac{p(S|h)p(h)}{p(S)}$, we can in fact maximize both $p(S \mid h)$ and $p(h)$, while ignoring the constant term $p(S)$. The following sections will expand upon this idea.

2 ML Hypothesis

The maximum likelihood (ML) hypothesis is defined when assuming that $p(h)$ is uniform: thus, no hypothesis is more likely than any other. As such, we have that

$$h_{\text{ML}} = \arg\max_h p(h \mid S) = \arg\max_h p(S \mid h)$$

One interesting observation that can be made is that if the data has no noise, the target concept is in the hypothesis class, and the hypothesis class is of finite size, then any hypothesis in the version space (i.e. hypotheses which are consistent with the data) is equally likely.

Another useful derivation is that, if we assume the underlying process that generates the

data has Gaussian noise with zero mean – that is, using as an example one dimensional data, for all $x_i \in \mathbb{R}$ and $y_i \in \mathbb{R}$ we have that

$$y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

for some target function f – it can be shown that

$$h_{\text{ML}} = \arg\max_{h \in \mathcal{H}} p(S \mid h) = \arg\min_{h \in \mathcal{H}} \sum_{i=1}^{|S|} (y_i - h(x_i))^2$$

In other words, the ML hypothesis is precisely the hypothesis that minimizes the sum of squared errors, no matter the hypothesis class. The only assumption is that of Gaussian noise. If other noise is present, the ML hypothesis would change.

3 MAP Hypothesis

The maximum *a posteriori* (MAP) hypothesis is defined when not making any assumptions about $p(h)$. Thus, we have that

$$h_{\text{MAP}} = \arg\max_{h \in \mathcal{H}} p(h \mid S) = \arg\max_{h \in \mathcal{H}} p(S \mid h) p(h)$$

From that, one can derive that

$$h_{\text{MAP}} = \arg\min_{h \in \mathcal{H}} \left(-\log_2 p(S \mid h) - \log_2 p(h) \right)$$

According to Information Theory, $-\log_2 p(S \mid h)$ and $-\log_2 p(h))$ represent the "lengths" of the data given the hypothesis and of the hypothesis, respectively. This leads to Minimum Description Length (MDL) view, which states that the best hypothesis is the one that minimizes both of them.

The length of the hypothesis can be thought of as relating to its complexity: more complex polynomials would require more bits to encode them, as would more complex decision trees, for example. The length of the data given the hypothesis can be thought of as a list of "corrections" that need to be made according to the predictions of the hypothesis: the less accurately the hypothesis predicts the data, the more "corrections" will need to be encoded (for example the difference between the hypothesis output and the observed $y_i \in \mathbb{R}$ in regression, or the correct label of the data point in classification).

As such, the MDL view states that the best hypothesis is the least complex which still describes the data well. In a sense, this can also be viewed as the trade-off between bias and variance.

4 Bayes Optimal Classifier

In the previous sections, finding the single best hypothesis under different assumptions was analyzed. However, it is possible to use all hypotheses in the hypothesis class to obtain

the Bayes Optimal Classifier for that class. The Bayes Optimal Classifier obtains the best average performance if compared to any hypothesis in the class. The output y^* returned by the Bayes Optimal Classifier for any data point is defined by

$$y^* = \arg\max_{y \in \mathcal{Y}} p(y \mid S) = \arg\max_{y \in \mathcal{Y}} \sum_h p(y \mid h) p(h \mid S)$$

In other words, it is the most likely label given the data, which is determined by the likelihood of the label for each hypothesis weighted by the likelihood of the hypothesis itself.