Lesson 9: Bayesian Learning

Bayesian learning

- Learn the **best** hypothesis given data and some domain knowledge.
- Learn the **most probable** h in hypothesis class H given data and domain knowledge. (**best** == **most probable**)

$$argmax_{h\ in\ H} Pr(h|D)$$

Bayes rule

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- $P(D)$ – prior on the data (**The shape of the world**)
- $P(h)$ – prior on h (this encapsulate our domain knowledge; **The rule of the world**)
- $P(D|h)$ – likelihood of data given the hypothesis. **Given the rule h is donimating the world, the probability that we can observe the world like D.**
- $P(h|D)$ – posteriori probability of hypothesis given data. **Given the observation D of the world, the probability of the rule h is ruling the world.**

Quiz: Bayes' rule

98% true positive. 2% false positive. 97% true negative. 3% false negative.

Only 0.8 percent of population has it.

$$P(T|P) = P(P|T) * P(T)/P(P) = .98 * .008/.0376 = 0.209$$

$$P(not\ T|P) = P(P|not\ T) * P(not\ T)/P(P) = .03 * .992/.0376 = 0.791$$

If 0.008 were higher, the test would be more useful.

Bayesian learning

For each h in H, calculate and get the max argument:

$$h_{MAP} = argmax_h P(h|D) \approx argmax_h P(D|h)P(h)$$

Maximum a posteriori (MAP). Note that the $P(D)$ is omitted and $\approx$ is used above.

$$h_{ML} = argmax_h P(D|h)$$

Maximum likelihood (ML): We further assume prior $P(h)$ is **uniform** so it is also dropped. (Just like we want to select the hypothesis that **best fits the data**.)

Direct computation not practical for large hypothesis spaces.

Bayesian learning in action

Assume: * We have labeled training data $\{< x_i, d_i >\}$ as noise-free examples of c. * c is in H * Uniform prior over hypothesis space.

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Uniform prior: $P(h) = 1/|H|$

For the posteriori:

$$P(D|h) = \begin{cases} 1 & if \;\; d_i = h(x_i) \;\; for \;\; all \;\; i \\ 0 & otherwise \end{cases} = \begin{cases} 1 & if \;\; h \;\; is \;\; in \;\; VS(D) \\ 0 & otherwise \end{cases}$$

Probability of the data:

$$\begin{aligned} P(D) &= \sum_i P(D|h_i)P(h_i) \quad (total\ probability) \\ &= \sum_{h\ in\ VS_H(D)} 1 * \frac{1}{|H|} = \frac{|VS|}{|H|} \end{aligned}$$

Put them into the equation, then for an h in the Version Space (And 0 for h outside):

$$P(h|D) = \frac{1 * 1/|H|}{|VS|/|H|} = \frac{1}{|VS|}$$

This means, given a bunch of data, the probability of a particular hypothesis being correct is uniform over elements of the version space.

Any element of the version space is good.

Bayesian learning with noise

Assume * We have labeled training data $\{< x_i, d_i >\}$ * $d_i = f(x_i) + e_i$ (where e_i is error) * $e_i \sim N(0, s^2)$ (i.i.d.)

Then the probability of data given the hypothesis is just the product of probabilities of the data points.

$$h_{ML} = argmax_h P(D|h) = argmax \prod_i P(d_i|h)$$

This distribution of the term $P(d_i|h)$ can be approximated with gaussian:

$$P(d_i|h) = \frac{1}{2\pi\sigma^2} * exp(-\frac{1}{2}\frac{(d_i - h(x_i))^2}{\sigma^2})$$

We can log it and ignore the constant terms for the purpose of argmax.

$$h_{ML} = argmax_h \sum_i -(d_i - h(x_i))^2 = argmin_h \sum_i (d_i - h(x_i))^2$$

This is the **sum of squared error**. . . . derived from the gaussian noise model and the maximum likelihood assumption. This means **sum of squared error** as a measure of how hypothesis fits the data has its background theoretical support.

But this also means that whenever you use sum of squared error, you are assuming the deterministic function $f(x_i)$ the i.i.d. Gaussian noise e_i in the data.

Bayesian learning

$$\begin{aligned} h_{MAP} &= argmax P(D|h)P(h) \\ &= argmax[lg(P(D|h)) + lg(P(h))] \\ &= argmin[-lg(P(D|h)) - lg(P(h))] \end{aligned}$$

In information theory, event w has probability p, i.e. has length $-lgp$. (lg here means log_2)

"Minimizing length($D|h$) + lenght(h)"

For the second term, for example, the smaller decision tree requires less description and thus we have bias on the smaller decision trees. It is the **size of** h. It is not only about the number of the parameters, but also about how big these parameters are.

If the hypothesis fits the data well, the data is superfluous information. Otherwise, we need a lot of data to describe the true data we got. So the first term correlates to **misclassifications errors**.

Minimum description length: It is a trade-off between the simplicity of the hypothesis and the low error of the hypothesis.

Bayesian classification

h_{MAP} and h_{ML} are how we get the best hypothesis. But when we want the best label for the data, we need to conduct a weighted vote.

Bayes Optimal Classifier

$$label_{MAP} = argmax_v \sum_h P(v|h)P(h|D)$$

(weighted vote of h in H based on $P(h|D)$)

Summary

- Bayes rule (swap "causes and effect")
  - $P(h|D) \sim P(D|h)P(h)$
- priors matter
- h_{MAP}, h_{ML}
- derived least square from Gaussian h_{ML}.
- best classification is consensus of all classifiers: **Bayes Optimal Classifier** (the best classifier you can possibly do)