Contents

0.1 Direct Learning

- Consider a distribution D
- X - Instance space, Y - Set of labels. (e.g. ± 1)
- Given a sample $\{(x, y)\}_1^n$ and a loss function $L(x, y)$, find a hypothesis

0.2 Probabilistic Model

Paradigm:

- Learn a probability distribution of the dataset.
- Use it to estimate which outcome is more likely.

Instead of learning $h : X \to Y$, learn $P(Y|X)$.

- Estimate probability from data
  - Maximum Likelihood Estimate (MLE)
  - Maximum Aposteriori Estimation (MAP)

0.3 Probability Recap

$$0 \leq P(A) \leq 1$$
$$P(true) = 1, P(false) = 0$$
$$P(A \vee B) = P(A) + P(B) + P(A \wedge B)$$
$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

0.4 Joint Distribution

Making a joint distribution of d variables

- Make a truth table listing all combinations of values of your variables (if there are d boolean variables then the table will have 2^d rows)
- For each combination of values, say how probable it is.
- The probability must sum up to 1.

Once we have the Joint Distribution, we find probability of any logical expression involving these variables.

$$P(E) = \sum_{rows\ matching\ E} P(row)$$

0.5 Independence

When two events do not affect each other's probabilities, they are called independent events

$$A \perp\!\!\!\perp B \leftrightarrow P(A \wedge B) = P(A) \times P(B)$$

The conditional independence of events A and B, given C is:

$$A \perp\!\!\!\perp B|c \leftrightarrow P(A|B,C) = \frac{P(A \wedge B|C)}{P(B|C)} = \frac{P(A|C) \times P(B|C)}{P(B|C)} = P(A|C)$$

0.6 Bayes' Rule

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \tag{1}$$

where A and B are events and $P(B) \neq 0$. Applying Bayes' rule for machine learning –

$$P(hypothesis \mid evidence) = \frac{P(evidence \mid hypothesis) \times P(hypothesis)}{P(evidence)} \tag{2}$$

0.7 Bayesian Learning

- Goal: find the best hypothesis from some space H of hypotheses, given the observed data (evidence) D.
- Define the most probable hypothesis in H to be the best.
- In order to do that, we need to assume a probability distribution over the class H.
- In addition, we need to know something about the relation ...

$P(h)$ – Prior Probability of the hypothesis h. Reflects the background knowledge, before data is observed.

$P(D)$ – Probability that this sample of the data is observed.

$P(D|h)$ – Probability of observing the sample D, given that hypothesis h is the target, also referred to as likelihood.

$P(h|D)$ – Posterior probability of h. The probability that h is the target, given that D has been observed.

- $P(h|D)$ increases with $P(h)$ and $P(D|h)$.
- $P(h|D)$ decreases with $P(D)$.

0.8 Maximum APosteriori Estimate

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

- The learner considers a set of candidate hypotheses H (models) and attempts to find the most probable one $h \in H$, given the observed data.
- Such maximally probable hypothesis is called maximum a posterior estimate (MAP). Bayes theorem is used to compute it:

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} P(h|D) \\ &= \arg\max_{h \in H} \frac{P(D|h) \times P(h)}{P(D)} \\ &= \arg\max_{h \in H} P(D|h) \times P(h) \end{aligned}$$

0.9 Maximum Likelihood Estimate

- We may assume that a priori, hypotheses are equally probable.

$$P(h_i) = P(h_j) \forall h_i, h_j \in H$$

- With that assumption, we can treat $\frac{P(h)}{P(D)}$ as a constant. We get the maximum likelihood estimate (MLE):

$$\begin{aligned} h_{MLE} &= \arg\max_{h \in H} \frac{P(D|h) \times P(h)}{P(D)} \\ &= \arg\max_{h \in H} P(D|h) \times P(h) \end{aligned}$$

- Here we just look for the hypothesis that best explains the data.

0.10 Bayesian Classifier

- $f : \vec{X} \to Y$ where, instances $x \in X$ is a collection of inputs –

$$\vec{x} = (x_1, x_2, \dots, x_n)$$

- Given an example, assign it the most probable value in Y.

$$\begin{aligned} y_{MAP} &= \arg\max_{y_j \in Y} P(y_j|x) \\ &= \arg\max_{y_j \in Y} P(y_j|x) \dots \end{aligned}$$

- Given the training data, we have to estimate the two terms.
- Estimating $P(y)$ is easy, e.g., under the binomial distribution assumption, count the number of times y appears in the training data.
- However, it is not feasible to estimate $P(x_1, x_2, \dots, x_n|y)$
- In this case, we have to estimate

0.11 Na ive Bayes Classifier

Assumption: Input feature values are independent, given the target value.

$$\begin{aligned} P(x_1, x_2, \dots, x_n|y_j) &= P(x_1|x_2, \dots, x_n, x_j) \times P(x_2, \dots, x_n|y) \\ &= P(x_1|x_2, \dots, x_n, x_j) \times P(x_2, \dots, x_n|y) \\ &= \vdots \\ &= \Pi_{i=1}^{n} P(x_i|y_j) \end{aligned}$$