Chapter 3

Naïve Bayes Learning

Contents

3.1 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

3.2 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)

3.3 Probability Recap

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

3.4 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.

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

3.5 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:

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

3.6 Bayes' Rule

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \tag{3.1}$$

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

$$P(hypothesis \mid evidence) = \frac{P(evidence \mid hypothesis) \times P(hypothesis)}{P(evidence)} \tag{3.2}$$

3.7 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 $P(D)$.

3.8 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:

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

3.9 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.

3.10 Bayesian Classifier

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

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

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

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

- 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 $P(x_1, x_2, \ldots, x_n|y)$
- In this case, we have to estimate

3.11 Na ive Bayes Classifier

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

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

3.12 Gaussian Naïve Bayes

Compute the mean and standard deviation to estimate the likelihood.

$$\mu_1 = E[X_1 \mid Y = 1] = \frac{2 + (-1.2) + 2.2}{3} = 1$$

$$\sigma_1^2 = E\left[(X_1 - \mu_1)^2 | Y = 1\right] = \frac{(2-1)^2 + (-1.2-1)^2 + (2.2-1)^2}{3} = 2.43$$

$$P(x_1 | Y = 1) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_1-\mu_1)^2}{2\sigma^2}} = \frac{1}{3.91} e^{-\frac{(x_1-1)^2}{4.86}}$$

3.13 Bayesian Belief Network

- Naïve Bayes classifier works with the assumption that the values of the input features are conditionally independent given the target value.
- This assumption dramatically reduces the complexity of learning the target function.
- Bayesian Belief Network describes the probability distribution governing a set of variables by specifying a set of conditional independence assumptions along with a set of conditional probabilities. Conditional independence assumptions here apply to subsets of the variables.

$$P(x_1, x_2, \ldots, x_l \mid x_1{}', x_2{}`, \ldots, x_m{}`, y_1, y_2, \ldots, y_n) = P(x_1, x_2, \ldots, x_l | y_1, y_2, \ldots, y_n)$$

3.14 Training Bayesian Classifier

During training, typically log-space is used.

$$\begin{aligned} y_{NB} &= \arg\max_y \left[\log P(y) \Pi_{i=1}^{n} P(x_i | y)\right] \\ &= \arg\max_y \left[\log P(y) + \sum_{i=1}^{n} \log P(x_i | y)\right] \end{aligned}$$

3.15 Text Classification

Algorithm 3.1 Text-based Naïve Bayes Classification

function TRAIN-NAIVE-BAYES(D, C) **returns** $\log P(c)$ and $\log P(w|c)$
    for all class $c \in C$ **do** ▷ Calculate $P(c)$ terms
        $N_{doc} \leftarrow$ number of documents in D
        $N_c \leftarrow$ number of documents from D in class c
        $logprior[c] \leftarrow \log \frac{N_c}{N_{doc}}$
        $V \leftarrow$ vocabulary of D
        $bigdoc[c] \leftarrow$ APPEND(d) **for** $d \in D$ **with** class c
        for all word w in V **do** ▷ Calculate $P(w|c)$ terms
            COUNT$(w, c) \ldots$
        end for
    end for
end function

The word with doesn't occur in the training set, so we drop it completely (we don't use unknown word models for Naïve Bayes)

3.16 Evaluating Classifiers

- Gold Label is the correct output class label of an input.
- Confusion Matrix is a table for visualizing how a classifier performs with respect to the fold labels, using two dimensions (system output and gold labels), and each cell labeling a set of possible outcomes.
- True Positives and True Negatives are correctly classified outputs belonging to the positive and negative class, respectively.

3.17 Precision, Recall, F-Measure

$$\textbf{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \tag{3.3}$$

3.18 ROC Curve

- A receiver operating characteristic curve (ROC curve) is a graphical plot that illustrates the performance of a binary classifier model.
- The ROC curve is the plot of the true positive rate (recall) (TPR) against the false positive rate (FPR).
- ROC curve plots TPR vs. FPR at different classification thresholds.

- Classification threshold is used to convert the output of a probabilistic classifier into class labels.
- The threshold determines the

3.19 Naïve Bayes: Two Classes

- Naïve Bayes classifier gives a method for predicting the most likely class rather than an explicit class.
- In the case of two classes, $y \in \{0, 1\}$ we predict that $y = 1$ iff

...

Take logarithm;

$$\log \frac{P(y_j = 1)}{P(y_j = 0)} + \sum_i \log \frac{1 - p_i}{1 - q_i} + \sum_i \left(\log \frac{p_i}{1 - p_i} - \log \frac{q_i}{1 - q_i} \right) x_i > 0$$

- We get that Naïve bayes is a linear separator with –

$$w_i = \log \frac{p_i}{1 - p_i} - \log \frac{q_i}{1 - q_i} = \log \frac{p_i(1 - q_i)}{q_i(1 - p_i)}$$

- In the case of two classes, we can say:
- but since $P(y_j = 1|x) = 1 - P(y_j = 0|x)$, we get:

$$P(y_j = 1|x) = \frac{1}{1 + e^{-(\sum_i w_i x_i + b)}}$$

- This is logistic regression