Discussion 04

2022.03.17

林维嘉 linwj@shanghaitech.edu.cn

Review

Estimating Probabilities

- Bayes Rule
- Maximum Likelihood Estimate (MLE)
- Maximum a Posterior (MAP)

Naïve Bayes

- Conditional Independence
- Naïve Bayes for Discrete Inputs
- Naïve Bayes for Continuous Inputs

Bayes Rule

Given observations D, our goal is to estimate the parameter θ. Through Bayes rule, we have the following identity,

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

where we call $P(\theta)$ the prior, $P(\theta|D)$ the posterior and $P(D|\theta)$ the likelihood.

MLE

One approach to estimate probabilities is to maximize the likelihood as follows,

$$\hat{\theta}^{MLE} = \arg\max_{\theta} P(D|\theta),$$

which is the general definition of MLE.

Intuition

We observe training data D, we should choose the value of θ that makes D most probable.

An Example

- X be a random variable for a coin, 1 or 0,
- θ is the probability of X taking 1, e.g.,$P(X = 1) = \theta$, and unknown,
- D is the observations produced by flip a coin X $N = \alpha_1 + \alpha_0$ times where α_1 the number of $X = 1$,
- Assuming I.I.D.

An Example

Likelihood is defined as $L(\theta) = P(D|\theta)$. With the conditions claimed before, we have the following formula,

$$L(\theta) = P(D|\theta) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}.$$

The MLE is to choose θ to maximize $P(D|\theta)$. For convenient, we take the log of $L(\theta)$,

$$l(\theta) = \ln L(\theta) = \alpha_1 \ln \theta + \alpha_0 \ln(1-\theta),$$

where $l(\theta)$ is called as log-likelihood. Since $l(\theta)$ is a concave function of θ, we just calculate the derivative of $l(\theta)$ with respect to θ:

An Example

$$\frac{\partial \ell(\theta)}{\partial \theta} = \frac{\partial \ln P(D|\theta)}{\partial \theta}$$

$$= \frac{\partial[\alpha_1 \ln \theta + \alpha_0 \ln(1 - \theta)]}{\partial \theta}$$

$$= \alpha_1 \frac{1}{\theta} + \alpha_0 \frac{-1}{(1 - \theta)} = 0$$

$$\rightarrow \quad \theta = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

Thus,

$$\hat{\theta}^{MLE} = \arg\max_{\theta} P(D|\theta) = \arg\max_{\theta} \ln P(D|\theta) = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

MAP

Given the observed data D and the prior $P(\theta)$, we want to maximize the posterior probability. By using Bayes rule, we have

$$\hat{\theta}^{MAP} = \arg\max_{\theta} P(\theta|D) = \arg\max_{\theta} \frac{P(D|\theta)P(\theta)}{P(D)}$$

Comparing the MLE algorithm, the only difference is the extra $P(\theta)$.

Intuition

Given new evidence, update the prior knowledge.

An Example

As in our coin flip example, the most common form of prior is a Beta distribution:

$$P(\theta) = Beta(\beta_0, \beta_1) = \frac{\theta^{\beta_1 - 1}(1 - \theta)^{\beta_0 - 1}}{B(\beta_0, \beta_1)}.$$

Recall the expression for $P(D|\theta)$, we have:

$$\begin{aligned}
\hat{\theta}^{MAP} &= \arg\max_{\theta} P(D|\theta)P(\theta) \\
&= \arg\max_{\theta} \theta^{\alpha_1}(1-\theta)^{\alpha_0} \frac{\theta^{\beta_1 - 1}(1 - \theta)^{\beta_0 - 1}}{B(\beta_0, \beta_1)} \\
&= \arg\max_{\theta} \frac{\theta^{\alpha_1 + \beta_1 - 1}(1 - \theta)^{\alpha_0 + \beta_0 - 1}}{B(\beta_0, \beta_1)} \\
&= \arg\max_{\theta} \theta^{\alpha_1 + \beta_1 - 1}(1 - \theta)^{\alpha_0 + \beta_0 - 1}
\end{aligned}$$

An Example

Substitute $(\alpha_1 + \beta_1 - 1)$ for α_1 and $(\alpha_0 + \beta_0 - 1)$ for α_0 in $\hat{\theta}^{MLE}$, we have

$$\hat{\theta}^{MAP} = \arg\max_{\theta} P(D|\theta)P(\theta) = \frac{(\alpha_1 + \beta_1 - 1)}{(\alpha_1 + \beta_1 - 1) + (\alpha_0 + \beta_0 - 1)}.$$

Eg. 2 Dice roll problem (6 outcomes instead of 2)

Likelihood is ~ Multinomial($\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$)

$$P(\mathcal{D} \mid \theta) = \theta_1^{\alpha_1}\theta_2^{\alpha_2} \ldots \theta_k^{\alpha_k}$$

If prior is Dirichlet distribution,

$$P(\theta) = \frac{\theta_1^{\beta_1 - 1}\ \theta_2^{\beta_2 - 1} \ldots \theta_k^{\beta_k - 1}}{B(\beta_1, \ldots, \beta_k)} \sim \text{Dirichlet}(\beta_1, \ldots, \beta_k)$$

Then posterior is Dirichlet distribution

$$P(\theta|D) \sim \text{Dirichlet}(\beta_1 + \alpha_1, \ldots, \beta_k + \alpha_k)$$

and MAP estimate is therefore

$$\hat{\theta}_i^{MAP} = \frac{\alpha_i + \beta_i - 1}{\sum_{j=1}^{k}(\alpha_j + \beta_j - 1)}$$

$$
\begin{aligned}
\ln P(\theta|\mathcal{D}) &\propto \ln\left(P(\mathcal{D}|\theta)P(\theta)\right) \\
&\propto \ln\left(\theta_1^{\alpha_1+\beta_1-1}\theta_2^{\alpha_2+\beta_2-1}\cdots\theta_K^{\alpha_K+\beta_K-1}\right) \\
&\propto \sum_{k=1}^{K}(\alpha_k+\beta_k-1)\ln\theta_k.
\end{aligned}
$$

Based on the fact that $\sum_{k=1}^{K}\theta_k = 1$, there are $K-1$ independent parameters in $\{\theta_1, \theta_2, ..., \theta_K\}$. Thus we can treat $\theta_K = 1 - \sum_{k=1}^{K-1}\theta_k$ as the dependent parameter. As the log-posterior is a concave function w.r.t. θ, its global maximum is obtained by setting its derivative equal to 0, leading to

$$
\begin{aligned}
\frac{\partial \ln P(\theta|\mathcal{D})}{\partial\theta_k} &= \frac{\alpha_k+\beta_k-1}{\theta_k} - \frac{\alpha_K+\beta_K-1}{1-\sum_{k=1}^{K-1}\theta_k} \\
&= \frac{\alpha_k+\beta_k-1}{\theta_k} - \frac{\alpha_K+\beta_K-1}{\theta_K} \\
&= 0.
\end{aligned}
$$

$$\hat{\theta}_k = \frac{\alpha_k + \beta_k - 1}{\alpha_K + \beta_K - 1}\hat{\theta}_K.$$

Substituting it into $\sum_{k=1}^{K} \theta_k = 1$,gives rise to

$$\hat{\theta}_K = \frac{\alpha_K + \beta_K - 1}{\sum_{k=1}^{K} \alpha_k + \beta_k - 1}.$$

$$\hat{\theta}_k = \frac{\alpha_k + \beta_k - 1}{\sum_{k=1}^{K} \alpha_k + \beta_k - 1}, \quad k = 1, 2, ..., K.$$

Conditional Independence

Definition: X is conditionally independent of Y given Z, if the probability distribution governing X is independent of the value of Y, given the value of Z

$$(\forall i, j, k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)$$

Which we often write

$$P(X|Y, Z) = P(X|Z)$$

Naïve Bayes assumes

$$P(X_1 \ldots X_n | Y) = \prod_i P(X_i | Y)$$

i.e., that X_i and X_j are conditionally independent given Y, for all $i \neq j$

Naïve Bayes for Discrete Inputs

We want to estimate two sets of parameters,

$$\theta_{ijk} \equiv P\left(X_i = x_{ij} \middle| Y = y_k\right),$$

and

$$\pi_k \equiv P(Y = y_k).$$

Naïve Bayes for Discrete Inputs

MLE:

$$\hat{\theta}_{ijk} = \hat{P}\left(X_i = x_{ij} \middle| Y = y_k\right) = \frac{\#D\{X_i = x_j \wedge Y = y_k\}}{\#D\{Y = y_k\}}$$

$$\hat{\pi}_k = P(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

MAP (Beta, Dirichlet priors):

$$\hat{\theta}_{ijk} = \hat{P}\left(X_i = x_{ij} \middle| Y = y_k\right) = \frac{\#D\{X_i = x_j \wedge Y = y_k\} + (\beta_k - 1)}{\#D\{Y = y_k\} + \sum_m (\beta_m - 1)}$$

$$\hat{\pi}_k = P(Y = y_k) = \frac{\#D\{Y = y_k\} + (\beta_k - 1)}{|D| + \sum_m (\beta_m - 1)}$$

Naïve Bayes for Continuous Inputs

Assume that for each possible discrete value y_k , the distribution of each continuous X_i is Gaussian. In order to train such a Naïve Bayes classifier we must therefore estimate the mean and standard deviation of each of these Gaussians:

$$\mu_{ik} = \boldsymbol{E}[X_i|Y = y_k],$$

$$\sigma_{ik}^2 = \boldsymbol{E}[(X_i - \mu_{ik})^2|Y = y_k].$$

Naïve Bayes for Continuous Inputs

MLE:

$$\hat{\mu}_{ik} = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j X_i^j \, \delta\left(Y^j = y_k\right),$$

$$\hat{\sigma}_{ik}^2 = \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j (X_i^j - \hat{\mu}_{ik})^2 \delta\left(Y^j = y_k\right).$$