Estimating Probabilities and Naive Bayes

Shuhao Xia

Shanghaitech University

March 30, 2020

Outline

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.

Continuing

Continuing

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,

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \frac{\partial \ln P(D|\boldsymbol{\theta})}{\partial \theta} \\ &= \frac{\partial \ln\left[\boldsymbol{\theta}^{\alpha_1}(1-\boldsymbol{\theta})^{\alpha_0}\right]}{\partial \theta}\end{aligned}$$

Continuing

$$
\begin{aligned}
&= \frac{\partial\left[\alpha_1 \ln\theta + \alpha_0 \ln(1-\theta)\right]}{\partial\theta} \\
&= \alpha_1 \frac{\partial \ln\theta}{\partial\theta} + \alpha_0 \frac{\partial \ln(1-\theta)}{\partial\theta} \\
&= \alpha_1 \frac{\partial \ln\theta}{\partial\theta} + \alpha_0 \frac{\partial \ln(1-\theta)}{\partial(1-\theta)} \cdot \frac{\partial(1-\theta)}{\partial\theta} \\
\implies \frac{\partial \ell(\theta)}{\partial\theta} &= \alpha_1 \frac{1}{\theta} + \alpha_0 \frac{1}{(1-\theta)} \cdot (-1) \\
\implies \theta &= \frac{\alpha_1}{\alpha_1 + \alpha_0}
\end{aligned}
$$

Thus,

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

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:

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

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

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

Continuing

Substitute $(\alpha_1 + \beta_1 - 1)$ for α_1 and $(\alpha_0 + \beta_0 - 0)$ for α_0 in $\hat{\theta}^{MLE}$, we have

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

Outline

Conditional Independence

Definition

$$P\left(Y=y_k|X_1\ldots X_n\right)=\frac{P\left(Y=y_k\right)\prod_i P\left(X_i|Y=y_k\right)}{\sum_j P\left(Y=y_j\right)\prod_i P\left(X_i|Y=y_j\right)}$$

The Naive Bayes classification rule:

$$\begin{aligned}Y&=\arg\max_{y_k}\frac{P\left(Y=y_k\right)\Pi_i P\left(X_i|Y=y_k\right)}{\sum_j P\left(Y=y_j\right)\Pi_i P\left(X_i|Y=y_j\right)}\\&=\arg\max_{y_k}P\left(Y=y_k\right)\prod_i P\left(X_i|Y=y_k\right).\end{aligned}$$

Naive Bayes for Discrete Input

We want to estimate two sets of parameters,

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

and

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

For example, by using MLE, we have

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

Adding a smoothing term, the estimate is given by

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

[U+FFFC]

Continuing

For π_k, we have

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

and the smoothed one

$$\hat{\pi}_k = \hat{P}\left(Y = y_k\right) = \frac{\#D\left\{Y = y_k\right\} + l}{|D| + lK}.$$

Naive 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 Naive Bayes classifier we must therefore estimate the mean and standard deviation of each of these Gaussians:

$$
\begin{aligned}
\mu_{ik} &= E\left[X_i | Y = y_k\right] \\
\sigma_{ik}^2 &= E\left[\left(X_i - \mu_{ik}\right)^2 | Y = y_k\right].
\end{aligned}
$$

By using MLE approach, we obtain:

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