Discussion on the Linearity of Naive Bayes

Lu Sun

March 27, 2020

Given input variables $X^\top = (X_1, X_2, ..., X_d) \in \mathbb{R}^d$ and corresponding output response $Y \in \{1, 2, ..., K\}$, we construct the naive Bayes classifier according to

$$P(Y|X) \propto \prod_{j=1}^{d} P(X_j|Y)P(Y). \tag{1}$$

Generally, naive Bayes is NOT a linear classifier. However, once the conditional probability $P(X_j|Y)$ $(\forall j)$ are from the *exponential family*

$$P(X_j|Y = k) = h(X_j)\exp(\boldsymbol{\eta}_{jk}^\top \boldsymbol{\phi}(X_j) - A(\boldsymbol{\eta}_{jk})), \quad \forall j, k, \tag{2}$$

the naive bayes classifier has a linear decision boundary in a *particular* feature space. In (2), $\boldsymbol{\eta}_{jk}$ denotes a parameter vector for X_j given $Y = k$, and $h(\cdot)$, $\phi(\cdot)$ and $A(\cdot)$ are know functions.

In naive Bayes, the decision boundary between classes k and ℓ $(\forall k \neq \ell)$ is determined by

$$\ln \frac{P(Y = k|X)}{P(Y = \ell|X)} = 0. \tag{3}$$

We rewrite the left hand of (3) as follows:

$$\begin{aligned}
\ln \frac{P(Y = k|X)}{P(Y = \ell|X)} &= \ln \frac{\prod_{j=1}^{d} P(X_j|Y = k)P(Y = k)}{\prod_{j=1}^{d} P(X_j|Y = \ell)P(Y = \ell)} \\
&= \sum_{j=1}^{d} \ln \frac{P(X_j|Y = k)}{P(X_j|Y = \ell)} + \ln \frac{P(Y = k)}{P(Y = \ell)} \\
&= \sum_{j=1}^{d} \big((\boldsymbol{\eta}_{jk} - \boldsymbol{\eta}_{j\ell})^\top \phi(X_j) - (A(\boldsymbol{\eta}_{jk}) - A(\boldsymbol{\eta}_{j\ell}))\big) + \ln \frac{\pi_k}{\pi_\ell} \\
&= \sum_{j=1}^{d} \boldsymbol{\beta}_j^\top \phi(X_j) + \beta_0,
\end{aligned} \tag{4}$$

where $\pi_k = P(Y = K)$ $(\forall k)$ and

$$\begin{aligned}
\boldsymbol{\beta}_j &= \boldsymbol{\eta}_{jk} - \boldsymbol{\eta}_{j\ell}, \\
\beta_0 &= \ln \frac{\pi_k}{\pi_\ell} - (A(\boldsymbol{\eta}_{jk}) - A(\boldsymbol{\eta}_{j\ell})).
\end{aligned} \tag{5}$$

Therefore, naive Bayes has a linear decision boundary in a transformed feature space $\phi(X)$, provided a exponential family $P(X_j|Y)$ $(\forall j)$.

In fact, a collection of probability distributions are from the exponential family, such as Bernoulli, categorical, binomial, multinomial, Gaussian, poisson, beta, Dirichlet, and so on. For example,

$$\begin{aligned}
\text{Bernoulli}: \quad & \phi(X_j) = X_j,\ X_j \in \{0,1\}, \\
\text{Categorical}: \quad & \phi(X_j) = \begin{pmatrix} \mathbf{1}_{X_j=1} \\ \vdots \\ \mathbf{1}_{X_j=M} \end{pmatrix},\ X_j \in \{1,2,...,M\}, \\
\text{Multinomial}: \quad & \phi(X_j) = X_j, \\
\text{Gaussian}: \quad & \phi(X_j) = \begin{pmatrix} X_j \\ X_j^2 \end{pmatrix},
\end{aligned} \tag{6}$$

where $\mathbf{1}_{X_j=m}$ is the indicator function, equaling 1 if $X_j = m$ and 0 otherwise, $\forall m$.

In conclusion, once $P(X_j|Y)$ follows Bernoulli or multinomial distributions, naive Bayes is a linear classifier. In contrast, once $P(X_j|Y)$ are from categorical or gaussian distributions, naive Bayes becomes a linear classifier in the transformed feature space $\phi(X)$.