나이브 베이즈와 그 응용

백승찬
서울대학교
베이즈 연구실

June 15, 2023

Table of contents

나이브 베이즈

1. 나이브 베이즈(Naive bayes)란? :

$$f_t(X_i) > f_j(X_i) \ \ \forall j \neq t, \qquad where \ \ f_j(X_i) = P(C_j) \prod_{k=1}^{p} P(X_{ik} = x_{ik}|C_j),$$

$$X_i = (X_{i1}, X_{i2}, \ldots, X_{ip}), \quad i = 1, \ldots, n, \quad C_j \in \{C_1, C_2, \ldots, C_k\}$$

$f(\mathbf{x}|y) = \prod_{k=1}^{p} f(x_k|y)$와 같이 조건부 독립성을 가정하고 확률을 추정하는 방식입니다. 이때 $\mathbf{x}$는 클래스 y를 결정짓는다고 판단되는 p개의 특성 $(x_1, x_2, \ldots, x_p)$로 구성됩니다.

가우시안 나이브 베이즈

1. 가우시안 나이브 베이즈(Gaussian naive bayes)란? : 주로 주어진 데이터셋이 범주형 값이 아닌 연속형 값을 가질때 쓰이는 방식이며, 이때 각각의 조건부 확률이 정규분포로 추정됩니다. 즉 다음과 같은 조건부 확률을 갖습니다 :

$$f(x_k|y) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x_k-\mu_k)^2}{2\sigma_k^2}}$$

   이때 μ_k와 σ_k^2는 y에 속한 x_k의 평균과 분산으로 추정할 수 있습니다.

2. 편의상 $y \in \{0, 1\}$이라고 하고 데이터 $\mathbf{x}$가 n개 있다고 하면 다음과 같은 데이터 행렬을 가정할 수 있습니다 :

$$\begin{bmatrix} y_1 & \mathbf{x}_1 \\ \vdots & \vdots \\ y_n & \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} y_1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \\ y_n & x_{n1} & \cdots & x_{np} \end{bmatrix}$$

   그러면 $\mu_k = \frac{1}{n}\sum_{i=1}^{n} x_{iO}$, $\sigma_k^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ik} - \mu_k)^2$, $k = 1, \ldots, p$으로 추정할 수 있습니다

가우시안 나이브 베이즈

3. 위와 같은 구조에서 가우시안 나이브 베이즈 분류기는 다음과 같이 구할 수 있습니다 :
Given traindataset $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, Classify x_{new} as $y = 1$ for

$$\left\{ \mathbf{x}_{new} : \prod_{k=1}^{p} \frac{1}{\sqrt{2\pi\sigma_{k_1}^2}} e^{-\frac{(\mathbf{x}_{new} - \mu_{k_1})^2}{2\sigma_{k_1}^2}} \pi(1) > \prod_{k=1}^{p} \frac{1}{\sqrt{2\pi\sigma_{k_0}^2}} e^{-\frac{(\mathbf{x}_{new} - \mu_{k_0})^2}{2\sigma_{k_0}^2}} \pi(0) \right\}$$

이때 μ_{k_j}와 $\sigma_{k_j}^2$, $j = 1, 2$는 다음과 같은 방식으로 추정할 수 있습니다 :

$$n_1 = \sum_{i=1}^{n} I(y_i = 1), \ \mu_{k_1} = \frac{1}{n_1} \sum_{i=1}^{n} x_{ik} I(y_i = 1), \ \sigma_{k_1}^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n} (x_{ik} - \mu_{k_1})^2 I(y_i = 1)$$

$$n_0 = \sum_{i=1}^{n} I(y_i = 0), \ \mu_{k_0} = \frac{1}{n_0} \sum_{i=1}^{n} x_{ik} I(y_i = 0), \ \sigma_{k_0}^2 = \frac{1}{n_0 - 1} \sum_{i=1}^{n} (x_{ik} - \mu_{k_0})^2 I(y_i = 0)$$

실제 데이터 분석

1. 다음과 같이 실제 축구데이터셋에 대해서 가우시안 나이브 베이즈를 적용해보았습니다 :
   - 2020-2021,2021-2022,2022-2023 시즌 유럽6대 축구리그 및 아시아6대 축구리그 실제 선수 데이터
   - 결측치 및 골키퍼 제외 총 13379개의 데이터로 이루어져있으며 분석에 불필요한 특성을 제외한 총 191개의 특성과 2가지의 클래스(아시아,유럽)으로 분류
   - 교차검증을 위해 2020-2021,2021-2022 시즌 데이터를 훈련데이터로, 2022-2023 시즌 데이터를 실험데이터로 분류
   - 훈련데이터는 총 8667개의 데이터, 실험데이터는 총 4712개의 데이터

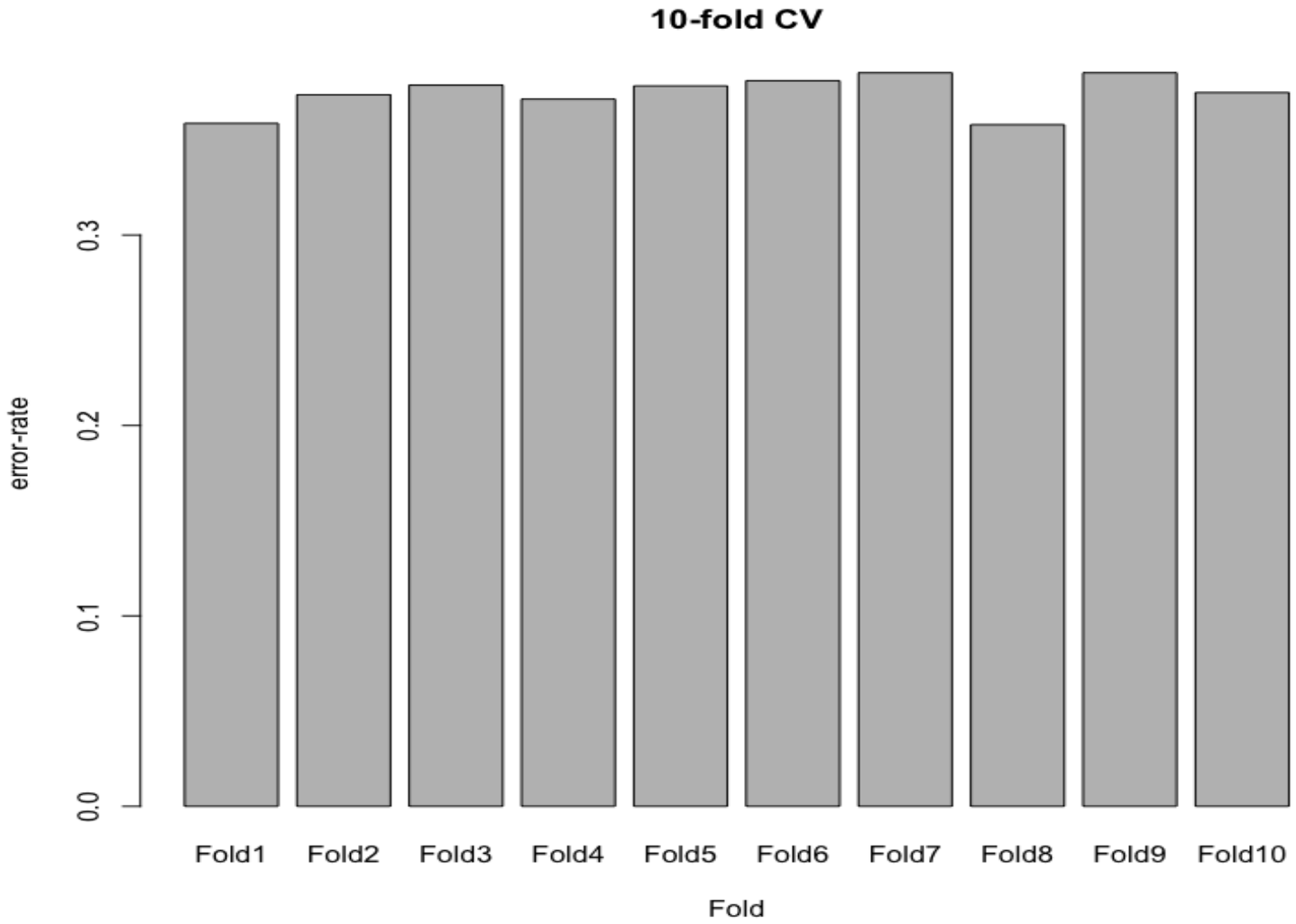
교차검증

2. 교차검증은 다음과 같은 방식으로 이루어졌습니다 :
데이터를 같은 크기의 K개 부분으로 나누고 k번째 부분에 관해서는 데이터의 다른 $K-1$ 부분으로 모델을 적합시키고, 데이터의 k번째 부분을 예측할 때는 적합된 모형의 예측오차를 계산합니다. 이를 $k = 1, 2, \ldots, K$에 관해 행하며 추정된 K개의 예측오차값의 평균을 구하면 이 값을 K-교차 검증 추정값이라고 할 수 있습니다. $\kappa : \{1, \ldots, N\} \to \{1, \ldots, K\}$를 관측치$i$가 확률화를 통해 어떠한 분할에 할당되는지를 가리키는 인덱스함수라고 하고 $\hat{f}^{-\kappa}(x)$를 적합된 함수라고 표기하면, 이는 데이터의 k번째 데이터를 삭제해 계산하게 됩니다. 그러면 다음과 같은 예측오차의 교차 검증 추정값을 사용할 수 있습니다 :

$$CV(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} I(y_i \neq \hat{f}^{-\kappa(i)}(x_i))$$

교차검증

3. 교차검증 결과는 다음과 같습니다 :



교차검증

4. 교차검증 결과는 다음과 같습니다 :

Fold	1	2	...	10
Error_rate	0.3587	0.3737	...	0.3748

즉 $CV(\hat{f}) = 0.3745$로 훈련데이터에 대해서 약 63프로의 정확도로 실제선수가 유럽리그에서 뛰는지를 구분할 수 있습니다.

실험결과비교

5. 실제 실험데이터에 대하여 가우시안 나이브 베이즈를 적용한 결과및 다른 방법론으로 적용했을때의 정확도에 관한 비교

방법론	Coin Toss	GNB	Decision Tree
정확도 (1-오류율)	0.5051	0.6464	0.8438

위 표에서 알 수 있듯이 가우시안 나이브 베이즈는 동전던지기보다는 좋은 성능을 보이나 절대적으로 좋은 성능을 가진다고는 할 수 없습니다. 그다지 높지 않은 정확도를 가지는 이유로는 다음과 같이 생각해볼 수 있습니다 :

- 모든 특성이 서로 독립적이라고 가정 $\Rightarrow$ 실제의 경우 이 가정은 참이 아닌 경우가 많으며, 특징들은 종속적인 관계를 가지는 경우가 있습니다.
- 데이터 부족 $\Rightarrow$ 일부 속성에 대해, 특정 클래스에 속하는 데이터가 적고 다른 클래스에 속하는 데이터가 많을 경우, 나이브 베이즈는 데이터가 더 많은 클래스에 편향됩니다.
- 무관한 특성에 대한 민감성 $\Rightarrow$ 나이브 베이즈는 무관한 특성에 대해 상당히 민감할 수 있습니다. 모든 특성을 고려하기 때문에 데이터셋에 많은 무관한 특성이 포함되어 있으면 모델 성능이 저하될 수 있습니다.

수정된 가우시안 나이브 베이즈

1. 앞서 설명드린 가우시안 나이브 베이즈 분류기와 비슷한 수정된 가우시안 나이브 베이즈 분류기 다음과 같이 구할 수 있습니다 :
   Given traindataset $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, Classify x_{new} as $y = 1$ for

$$\left\{ \mathbf{x}_{new} : \prod_{k=1}^{p} \frac{1}{\sqrt{2\pi\hat{\sigma}_{k_1}^2}} e^{-\frac{(\mathbf{x}_{new} - \hat{\mu}_{k_1})^2}{2\hat{\sigma}_{k_0}^2}} \pi(1) > \prod_{k=1}^{p} \frac{1}{\sqrt{2\pi\hat{\sigma}_{k_0}^2}} e^{-\frac{(\mathbf{x}_{new} - \hat{\mu}_{k_0})^2}{2\hat{\sigma}_{k_0}^2}} \pi(0) \right\}$$

2. 이때 μ_{k_j}와 $\sigma^2_{k_j}$, $j = 1, 2$는 다음과 같은 방식으로 추정할 수 있습니다 :

$$\mu_{k_1}^{(-j)} = \frac{1}{n_1 - 1}\sum_{i=1}^{n} x_{ik} I(i \neq j, y_i = 1),\ \hat{\mu}_{k_1} = \frac{1}{n_1}\sum_{j=1}^{n} \mu_{k_1}^{(-j)}, \tag{1}$$

$$\mu_{k_0}^{(-j)} = \frac{1}{n_0 - 1}\sum_{i=1}^{n} x_{ik} I(i \neq j, y_i = 0),\ \hat{\mu}_{k_0} = \frac{1}{n_0}\sum_{j=1}^{n} \mu_{k_0}^{(-j)}, \tag{2}$$

$$\sigma_{k_1}^{2(-j)} = \frac{1}{n_1 - 2}\sum_{i=1}^{n} (x_{ik} - \mu_{k_1}^{(-j)})^2 I(i \neq j, y_i = 1), \tag{3}$$

$$\sigma_{k_0}^{2(-j)} = \frac{1}{n_0 - 2}\sum_{i=1}^{n} (x_{ik} - \mu_{k_0}^{(-j)})^2 I(i \neq j, y_i = 0) \tag{4}$$

$$\therefore \hat{\mu}_{k_1} = \frac{1}{n_1}\sum_{j=1}^{n} \mu_{k_1}^{(-j)},\ \hat{\mu}_{k_0} = \frac{1}{n_0}\sum_{j=1}^{n} \mu_{k_0}^{(-j)},\ \hat{\sigma}^2_{k_1} = \frac{1}{n_1}\sum_{j=1}^{n_1} \sigma_{k_1}^{2(-j)},\ \hat{\sigma}^2_{k_0} = \frac{1}{n_0}\sum_{j=1}^{n_0} \sigma_{k_0}^{2(-j)}$$

Future work

1. 각 특성에 대한 가중치를 고려한 방법론
2. 배깅을 적용한 가우시안 나이브 베이즈 방법론
3. 수정된 가우시안 나이브 베이즈 방법론

Reference

P. Domingos, M. Pazzani (1996). Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier. ICML

P. Domingos, M. Pazzani (1997). on the optimality of the simple bayesian classifier under zero-one loss. Machine Learning

T.Hastie, R.Tibshirani, J.Friedman (2017). The Elements of Statistical Learning. Springer

감사합니다