Naive Bayes & regressão logística

Henrique Laureano
http://leg.ufpr.br/~henrique

CiDWeek I, 03-07/02/2020

UNIVERSIDADE FEDERAL DO PARANÁ

Naive Bayes

Primeiro, precisamos falar sobre o que é um classificador de Bayes.

Classificador de Bayes

Um *framework* probabilístico para problemas de classificação, baseado no teorema de Bayes.

Exemplo,

- Meningite causa torcicolo 50% das vezes, $\mathbb{P}[T|M]$
- Prob. *a priori* de um paciente estar com meningite é $1/50.000$, $\mathbb{P}[M]$
- Probabilidade *a priori* de um paciente estar com torcicolo é $1/20$, $\mathbb{P}[T]$

Se um paciente está com torcicolo, qual a probabilidade dele estar com meningite?

$$\mathbb{P}[M|T] = \frac{\mathbb{P}[T|M]\ \mathbb{P}[M]}{\mathbb{P}[T]} = \frac{1/2 \times 1/50.000}{1/20} = 0.0002.$$

Classificadores Bayesianos

Considere atributos $A_1, A_2, \dots A_n$ e uma classe C com rótulos $c_1, c_2, \dots c_k$.

O que queremos?

$$\text{Predição}: \quad C = c_1 \text{ ou } C = c_2 \text{ ou } \dots,$$

i.e., queremos o valor de C que maximiza $\mathbb{P}[C|A_1, A_2, \dots, A_n]$.

Como fazemos? Teorema de Bayes.

Calculamos a probabilidade *a posteriori* $\mathbb{P}[C|A_1, A_2, \dots, A_n]$ para todos os valores de C,

$$\mathbb{P}[C|A_1, A_2, \dots, A_n] = \frac{\mathbb{P}[A_1, A_2, \dots, A_n|C] \; \mathbb{P}[C]}{\mathbb{P}[A_1, A_2, \dots, A_n]}.$$

E como calculamos $\mathbb{P}[A_1, A_2, \dots, A_n|C]$? Naive Bayes.

CiDAMO

Classificador Naive Bayes

Por que *naive*?

Porque se assume independência entre os atributos A_i, i.e.,

$$\mathbb{P}[A_1, A_2, \ldots, A_n | C_k] = \mathbb{P}[A_1|C_k]\ \mathbb{P}[A_2|C_k]\ \ldots\ \mathbb{P}[A_n|C_k].$$

Vantagem: Grande redução do custo computational.

Um novo ponto é classificado como C_j se $\mathbb{P}[C_j] \times \prod_{i=1}^{n} \mathbb{P}[A_i|C_k]$ é máximo.

CiDAMO

Exemplo: Estimando probabilidades a partir dos dados

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- $\mathbb{P}[C] = N_k/N$
  - $\mathbb{P}[C = \text{No}] = 7/10$
  - $\mathbb{P}[C = \text{Yes}] = 3/10$

Atributos discretos:

- $\mathbb{P}[A_i|C_k] = A_{ik}/N_k$
  - $\mathbb{P}[\text{Status} = \text{Married}|\text{No}] = 4/7$
  - $\mathbb{P}[\text{Refund} = \text{Yes}|\text{Yes}] = 0$
  - ...

CiDAMO

E com atributos contínuos?

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Estimação da densidade de probabilidade

- Se assume distribuição Normal
- Se estima a média μ e o desvio padrão σ
- Se estima a probabilidade condicional

$$\mathbb{P}[A_i | C_k] = \frac{\exp\left\{-\frac{(A_i - \mu_{ik})^2}{2\sigma_{ik}^2}\right\}}{\sqrt{2\pi\sigma_{ik}^2}}$$

Exemplo,

$$\mathbb{P}[\text{Income} = 120|\text{No}] = \frac{1}{\sqrt{2\pi 2975}} \exp\left\{-\frac{(120 - 110)^2}{22975}\right\}$$
$$= 0.0072.$$

CiDAMO

Classificador Naive Bayes: Exemplo

Dado o perfil: $X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{k})$

$$
\begin{aligned}
\mathbb{P}[X|\text{Class} = \text{No}] &= \mathbb{P}[\text{Refund} = \text{No}|\text{Class} = \text{No}] \times \\
&\quad \mathbb{P}[\text{Married}|\text{Class} = \text{No}] \times \\
&\quad \mathbb{P}[\text{Income} = 120k|\text{Class} = \text{No}] \\
&= 4/7 \times 4/7 \times 0.0072 = 0.0024.
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{P}[X|\text{Class} = \text{Yes}] &= \mathbb{P}[\text{Refund} = \text{No}|\text{Class} = \text{Yes}] \times \\
&\quad \mathbb{P}[\text{Married}|\text{Class} = \text{Yes}] \times \\
&\quad \mathbb{P}[\text{Income} = 120k|\text{Class} = \text{Yes}] \\
&= 1 \times 0 \times 10^{-9} = 0.
\end{aligned}
$$

Já que $\mathbb{P}[X|\text{No}]\ \mathbb{P}[\text{No}] > \mathbb{P}[X|\text{Yes}]\ \mathbb{P}[\text{Yes}]$,

$$\Rightarrow \quad \mathbb{P}[X|\text{No}] > \mathbb{P}[X|\text{Yes}] \Rightarrow \text{Class} = \text{No}.$$

"Dibrando" o problema de probabilidade zero

Going a little deeper in the smoothing penalty

Smoothing penalty leads to an optimal curve, the smoothing spline[1]. The penalty for smoothing splines takes the form $J(\beta, \lambda) = \lambda \int (Df)^2 = \lambda \langle Df, Df \rangle$.

$$\text{When } f(x) = \sum_{j=1}^{M} \beta_j \psi_j(x), \text{ we have } J(\beta, \lambda) = \lambda \beta^\top \boldsymbol{S} \beta$$

where $\boldsymbol{S}$ is a $M \times M$ matrix with $(i, j)^{\text{th}}$ entry $\langle D\psi_i, D\psi_j \rangle$.

Rewriting the penalized log-likelihood as a likelihood,

$$\exp\{l_p(\beta, \lambda)\} = \exp\{l(\beta)\} \times \exp(-\lambda \beta^\top \boldsymbol{S} \beta),$$

$\exp(-\lambda \beta^\top \boldsymbol{S} \beta)$ is $\propto$ to a $\text{MVN}(0, \boldsymbol{S}_\lambda^{-1} = (\lambda \boldsymbol{S})^{-1})$.
The penalized likelihood is equivalent to assigning the prior $\beta \sim \text{MVN}(0, \boldsymbol{S}_\lambda^{-1})$.

CiDAMO

Connection: SPDE model as a basis-penalty smoother

- For a given differential operator D, the approx. $\boldsymbol{Q}$ for the SPDE is the same as the precision matrix $\boldsymbol{S}_\lambda$ computed using the smoothing penalty $\langle Df, Df \rangle$;
- Differences between the basis-penalty approach and the SPDE finite element approx., when using the same basis and differential operator, are differences in implementation only.

Lindgren, F., Rue, H. and Lindström, J. (2011)[a]

[a] An Explicit Link between Gaussian Fields and Gaussian Markov Random Fields: The Stochastic Partial Differential Equation Approach (with discussion). *Journal of the Royal Statistical Society: Series B 73*(4), 423-498

An approx. solution to the SPDE is given by representing f as a sum of linear (specifically, B-spline) basis functions multiplied by coefficients; the coefs of these basis form a GMRF.

Matérn penalty

$$D = \tau(\kappa^2 - \Delta) \quad \Rightarrow \quad \text{smoothing penalty} : \tau \int (\kappa^2 f - \Delta f)^2 \, dx.$$

- inverse correlation range κ: higher values lead to less smooth functions;
- smoothing parameter τ controls the overall smoothness of f.

In matrix form, this leads to the smoothing matrix

$$\boldsymbol{S} = \tau(\kappa^4 \boldsymbol{C} + 2\kappa^2 \boldsymbol{G}_1 + \boldsymbol{G}_2) \quad \text{where}$$

$\boldsymbol{C}, \boldsymbol{G}_1, \boldsymbol{G}_2$ are all $M \times M$ sparse matrices with $(i,j)^{\text{th}}$ entries $\langle \psi_i, \psi_j \rangle, \langle \psi_i, \nabla \psi_j \rangle$, and $\langle \nabla \psi_i, \nabla \psi_j \rangle$.

The matrix $\boldsymbol{S}$ is equal to the matrix $\boldsymbol{Q} = \boldsymbol{P}^\top \boldsymbol{Q}_e \boldsymbol{P}$ computed using the FEM.

Fitting the Matérn SPDE in `mgcv`

`mgcv` allows the specification of new basis-penalty smoothers.

step-by-step

- `INLA::inla.mesh.(1d or 2d)` to create a mesh;
- `INLA::inla.mesh.fem` to calculate $\boldsymbol{C}$, $\boldsymbol{G}_1$, and $\boldsymbol{G}_2$;
- Connect the basis representation of f to the observation locations,
  - The full design matrix is given by combining the fixed effects design matrix $\boldsymbol{X}_c$ and the contribution for f, $\boldsymbol{A}$ - the projection matrix found using `INLA::inla.spde.mesh.A`;
- Use REML to findo optimal κ, τ and β.

Some final remarks,

- As REML is an empirical Bayes procedure, we expect point estimates for $\hat{\beta}$ to coincide with `R-INLA`;
- A uniform prior is implied for the smoothing parameters τ and κ;
- `R-INLA` allows for similar estimation by just using the modes of the hyperparameters κ and τ (`int.strategy="eb"`).

To finish, let's check some [code].

CiDAMO