Optimal decision and naive Bayes

Fabrice Rossi

SAMM
Université Paris 1 Panthéon Sorbonne

2017

Outline

Standard supervised learning hypothesis

Optimal model

The Naive Bayes Classifier

Data model

Data spaces

- $\mathcal{X}$: "input" space, always observed
- $\mathcal{Y}$: "output" space, values to predict, observed during learning

Minimal structural assumption

$\mathcal{X}$ should be equipped with a dissimilarity d:

- d is a function from $\mathcal{X} \times \mathcal{X}$ to $\mathbb{R}^+$
- d is symmetric
- $\forall \mathbf{X}, \mathbf{X}' \in \mathcal{X}, \quad d(\mathbf{X}, \mathbf{X}) \leq d(\mathbf{X}, \mathbf{X}')$

Multivariate assumption

In general $\mathcal{X}$ is structured into "variables":

- $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_P$
- $\mathbf{X} = (X_1, X_2, \ldots, X_P)^T$

Predictive model

Definition and use

- A predictive model is a function g from $\mathcal{X}$ to $\mathcal{Y}$
- for an observation $(\mathbf{X}, \mathbf{Y})$ one expects $g(\mathbf{X})$ to be "close to" $\mathbf{Y}$

Loss function

A loss function l is

- a function from $\mathcal{Y} \times \mathcal{Y}$ to $\mathbb{R}^+$
- such that $\forall \mathbf{Y} \in \mathcal{Y}, \quad l(\mathbf{Y}, \mathbf{Y}) = 0$

Interpretation

$l(g(\mathbf{X}), \mathbf{Y})$ measures the loss incurred by the user of a model g when the true value $\mathbf{Y}$ is replaced by the prediction $g(\mathbf{X})$.

Examples

$\mathcal{Y} = \mathbb{R}$

- $l_2(p, t) = (p - t)^2$
- $l_1(p, t) = |p - t|$
- $l_{APE}(p, t) = \frac{|p-t|}{|t|}$

$|\mathcal{Y}| < \infty$

- $l_b(p, t) = \mathbf{1}_{p \neq t}$
- general case when $\mathcal{Y} = \{y_1, y_2\}$

$l(p, t)$	$t = y_1$	$t = y_2$
$p = y_1$	0	$l(y_1, y_2)$
$p = y_2$	$l(y_2, y_1)$	0

asymmetric costs are important in practice (think SPAM versus non SPAM)

Stochastic framework

Main hypotheses

- observations are random variables with values in $\mathcal{X} \times \mathcal{Y}$
- they are distributed according to a fixed and unknown distribution D
- observations are independent

A data set

- $\mathcal{D} = ((\mathbf{X}_i, \mathbf{Y}_i))_{1 \leq i \leq N}$
- $(\mathbf{X}_i, \mathbf{Y}_i) \sim D$ and $\mathcal{D} \sim D^N$
- notation: $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{iP})^T$

Risk of a model

The risk of g for the loss function l is

$$R_l(g) = \mathbb{E}_{(\mathbf{X},\mathbf{Y})\sim D}(l(g(\mathbf{X}), \mathbf{Y}))$$

Technical aspects

Additional hypotheses

- there is an underlying probability space
- $\mathcal{X} \times \mathcal{Y}$ must be a measurable space
- in general this is done via the standard Borel sigma field on $\mathbb{R}^d$

Measurability

- loss functions must be measurable functions
- ditto for models
- technically, the loss could be $+\infty$

Independence and stationarity

- independence can be relaxed for e.g. time series
- stationrity also for e.g. drift analysis

Outline

Standard supervised learning hypothesis

Optimal model

The Naive Bayes Classifier

Optimal model

Optimal risk

- $R_l^* = \inf_g R_l(g)$
- called the *Bayes risk* when $\mathcal{Y}$ is finite

Is R_l^* reachable?

- in general $\arg\min_g R_l(g)$ is a set: could it be empty?
- g_l^*: a model such that $R_l(g_l^*) = R_l^*$

Optimal model

Optimal risk

- $R_l^* = \inf_g R_l(g)$
- called the *Bayes risk* when $\mathcal{Y}$ is finite

Is R_l^* reachable?

- in general $\arg\min_g R_l(g)$ is a set: could it be empty?
- g_l^*: a model such that $R_l(g_l^*) = R_l^*$

Quadratic case

- $\mathcal{Y} = \mathbb{R}$, $l_2(p, v) = (p - v)^2$
- $g^*(\mathbf{x}) = \mathbb{E}_{(\mathbf{X},Y)\sim D}(Y|\mathbf{X} = \mathbf{x})$

Discrete case

If $\mathcal{Y}$ is finite

- a loss function is a table with $|\mathcal{Y}|(|\mathcal{Y}|-1)$ non zero entries
- g^* can be obtained using conditional probabilities $\mathbb{P}(Y=y|\mathbf{X}=\mathbf{x})$

Simple case

- $l_b(p,t)=\mathbf{1}_{p\neq t}$
- $g^*_{l_b}(\mathbf{x})=\arg\max_{y\in\mathcal{Y}}\mathbb{P}_{(\mathbf{X},\mathbf{Y})\sim D}(Y=y|\mathbf{X}=\mathbf{x})$

General case

$$g^*_l(\mathbf{x})=\arg\min_{y\in\mathcal{Y}}\sum_{y'\neq y} l(y,y')\mathbb{P}_{(\mathbf{X},Y)\sim D}(Y=y'|\mathbf{X}=\mathbf{x})$$

Idea of the proof

- conditional reasoning

$$R_l(g) = \mathbb{E}_{(\mathbf{X},Y)\sim D}\left\{\mathbb{E}_{(\mathbf{X},Y)\sim D}(l(g(\mathbf{x}), Y)|\mathbf{X} = \mathbf{x})\right\}$$

- standard properties of the expectation

$$\mathbb{E}_{(\mathbf{X},Y)\sim D}(l(g(\mathbf{x}), Y)|\mathbf{X} = \mathbf{x}) = \sum_{y'\in\mathcal{Y}} l(g(\mathbf{x}), y')\mathbb{P}_{(\mathbf{X},Y)\sim D}(Y = y'|\mathbf{X} = \mathbf{x})$$

- pointwise minimization and $l(y, y) = 0$ gives the result

Remarks

Discriminant versus Generative

- Bayes rule: $\mathbb{P}(Y = y|\mathbf{X} = \mathbf{x}) = \frac{\mathbb{P}(\mathbf{X} = \mathbf{x}|Y = y)\mathbb{P}(Y = y)}{\mathbb{P}(\mathbf{X} = \mathbf{x})}$
- for a fixed $\mathbf{x}$, one can compare the $\mathbb{P}(\mathbf{X} = \mathbf{x}|Y = y)\mathbb{P}(Y = y)$ rather than the $\mathbb{P}(Y = y|\mathbf{X} = \mathbf{x})$
- Y given $\mathbf{X}$: discriminant, $\mathbf{X}$ given Y: generative

$\mathcal{Y} = \{y_1, y_2\}$

$$g_l^*(\mathbf{x}) = \begin{cases} y_1 & \text{if } \frac{l(y_1, y_2)\mathbb{P}(Y = y_2|\mathbf{X} = \mathbf{x})}{l(y_2, y_1)\mathbb{P}(Y = y_1|\mathbf{X} = \mathbf{x})} \leq 1 \\ y_2 & \text{in the other case} \end{cases}$$

Ratios of probabilities are sufficient to compute the optimal model

Outline

Standard supervised learning hypothesis

Optimal model

The Naive Bayes Classifier

Generative models

Generative model

- a model that explains both **X** and **Y**
- as opposed to **Y** given **X**
- one road to optimal models

Hypotheses

- $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_P$
- $\mathcal{Y}$ and all the $\mathcal{X}_j$ are finite sets

Probability distributions on $\mathcal{X} \times \mathcal{Y}$

- $|\mathcal{X}_1| \times |\mathcal{X}_2| \times \ldots \times |\mathcal{X}_P| \times |\mathcal{Y}| - 1$ parameters
- generally intractable

A simple model

Conditional independence

- explanatory variables are assumed independent given the target variable
- $\perp\!\!\!\perp_{1 \leq j \leq P} X_j | \mathbf{Y}$ and thus

$$\mathbb{P}(\mathbf{X} = \mathbf{x} | Y = y) = \prod_{j=1}^{P} \mathbb{P}(X_j = x_j | Y = y)$$

Consequences

- only $\left(1 + \sum_{j=1}^{P} (|\mathcal{X}_j| - 1)\right) |\mathcal{Y}| - 1$ parameters
- easy estimation
- but very strong assumption

Naive Bayes

Categorical distribution

- arbitrary distribution on $\mathcal{X}_j = \left\{u_1^{(j)}, \ldots, u_{|\mathcal{X}_j|}^{(j)}\right\}$
- parameter vector $\Gamma = (\gamma_1, \ldots, \gamma_{|\mathcal{X}_j|})$
- $\mathbb{P}_{X \sim C(\Gamma)}(X = u_l^{(j)}) = \gamma_l$ and by extension $\mathbb{P}_{X \sim C(\Gamma)}(X = u) = \gamma_u$

Naive Bayes distribution

- $\mathbf{\Gamma} = (\Gamma_Y, (\Gamma_{j,y})_{1 \leq j \leq P, y \in \mathcal{Y}})$
- distribution on $\mathcal{X} \times \mathcal{Y}$

$$\mathbb{P}_{(\mathbf{X},\mathbf{Y}) \sim NB(\mathbf{\Gamma})}(\mathbf{X} = \mathbf{x}, Y = y) = \mathbb{P}_{Y \sim C(\Gamma_Y)}(Y = y) \prod_{j=1}^{P} \mathbb{P}_{X_j | Y = y \sim C(\Gamma_{j,y})}(X_j = x_j | Y = y)$$

Maximum Likelihood Estimation

MLE

- $\widehat{\mathbf{\Gamma}}_{MLE} = \arg\max_{\mathbf{\Gamma}} \mathbb{P}(\mathcal{D}|\mathbf{\Gamma})$
- equivalently maximizing $\log \mathbb{P}(\mathcal{D}|\mathbf{\Gamma})$

Naive Bayes log Likelihood

- standard i.i.d. assumptions
- separability

$$
\begin{aligned}
\log \mathbb{P}(\mathcal{D}|\mathbf{\Gamma}) &= \sum_{i=1}^{N} \log \mathbb{P}_{(\mathbf{X},\mathbf{Y})\sim NB(\mathbf{\Gamma})}(\mathbf{X} = \mathbf{X}_i, Y = Y_i) \\
&= \sum_{i=1}^{N} \sum_{j=1}^{P} \log \mathbb{P}_{X_j|Y=y_i\sim C(\Gamma_{j,y})}(X_j = X_{ij}|Y = Y_i) \\
&\quad + \sum_{i=1}^{N} \log \mathbb{P}_{Y\sim C(\Gamma_Y)}(Y = Y_i)
\end{aligned}
$$

MLE cont.

Separate estimations

- parameters of the different categorical distributions are independent
- $\widehat{\boldsymbol{\Gamma}}_{MLE}$ can be computed block by block

$$\widehat{\Gamma_Y}_{MLE} = \arg\max_{\Gamma_Y} \sum_{i=1}^{N} \log \mathbb{P}_{Y\sim C(\Gamma_Y)}(Y = Y_i)$$

$$\widehat{\Gamma_{j,y}}_{MLE} = \arg\max_{\Gamma_{j,y}} \sum_{i=1, Y_i=y}^{N} \log \mathbb{P}_{X_j|Y=y\sim C(\Gamma_{j,y})}(X_j = X_{ij}|Y = y)$$

Consequences

- simple unidimensional estimation
- conditional frequencies

Frequencies

Target variable

$$\forall y \in \mathcal{Y}, \widehat{\gamma_{Y,y}}_{MLE} = \frac{|\{i | Y_i = y\}|}{N}$$

Explanatory variables

$$\forall y \in \mathcal{Y}, \forall j, \forall x \in \mathcal{X}_j, \widehat{\gamma_{j,y,x}}_{MLE} = \frac{|\{i | Y_i = y \text{ and } X_{i,j} = x\}|}{|\{i | Y_i = y\}|}$$

Computational cost

Very efficient method: $O(NP)$ (with a one pass algorithm)

Extensions

Infinite discrete set

- $\mathcal{X}_j = \mathbb{N}$
- replace the categorical distribution by e.g. a Poisson distribution (or any distribution on $\mathbb{N}$)
- frequency based estimation, e.g. if $X_j | Y = y$ is a Poisson distribution with parameter $\lambda_{j,y}$ then

$$\widehat{\lambda_{j,y}}_{MLE} = \frac{\sum_{i, Y_i = y} X_{i,j}}{|\{i | Y_i = y\}|}$$

Continuous variables

- $\mathcal{X}_j = \mathbb{R}$
- same principle: choose a parametric distribution on $\mathbb{R}$, e.g. the Gaussian distribution
- block based estimation

Naive Bayes Classifier

Strategy

- estimate the parameters of a NB model $NB(\mathbf{\Gamma})$ for a data set
- approximate the data distribution by the *NB* distribution i.e.

$$\mathbb{P}_{(\mathbf{X},Y)\sim D}(\mathbf{x},y) \simeq \mathbb{P}_{(\mathbf{X},Y)\sim NB(\widehat{\mathbf{\Gamma}}_{MLE})}(\mathbf{x},y)$$

- use the approximation to compute the “optimal” classifier, i.e.

$$g_l^*(\mathbf{x}) \simeq \arg\min_{y\in\mathcal{Y}} \sum_{y'\neq y} l(y,y')\mathbb{P}_{(\mathbf{X},Y)\sim NB(\widehat{\mathbf{\Gamma}}_{MLE})}(Y=y'|\mathbf{X}=\mathbf{x})$$

- the classifier is optimal only for data distributed exactly according to $NB(\widehat{\mathbf{\Gamma}}_{MLE})$: this is seldom the case in practice!

Naive Bayes Classifier

Example

- as $\mathbf{x}$ is fixed when one computes $g_l^*(\mathbf{x})$, only $\mathbb{P}_{(\mathbf{X},Y)\sim NB(\widehat{\mathbf{\Gamma}}_{MLE})}(Y=y',\mathbf{X}=\mathbf{x})$ is needed
- we have

$$\mathbb{P}_{(\mathbf{X},Y)\sim NB(\widehat{\mathbf{\Gamma}}_{MLE})}(Y=y,\mathbf{X}=\mathbf{x}) = \mathbb{P}_{Y\sim C(\widehat{\Gamma_{Y}}_{MLE})}(Y=y)\prod_{j=1}^{P}\mathbb{P}_{X_j|Y=y\sim C(\widehat{\Gamma_{j,y}}_{MLE})}(X_j=x_j|Y=y)$$

and thus

$$\mathbb{P}_{(\mathbf{X},Y)\sim NB(\widehat{\mathbf{\Gamma}}_{MLE})}(Y=y,\mathbf{X}=\mathbf{x}) = \widehat{\gamma_{Y,y}}_{MLE}\prod_{j=1}^{P}\widehat{\gamma_{j,y,\mathbf{x}_j}}_{MLE}$$

- very simple frequency comparison!

In practice

Pros and Cons

+ very fast
+ handles mixed data easily
- limited predictive performances compared to state-of-the-art methods
- needs a very good set of explanatory variables

Best practices

- avoid using the NBC when frequencies are very close to 0 or 1
- use a variable selection method

Licence

Version

Last git commit: 2018-06-06
By: Fabrice Rossi (Fabrice.Rossi@apiacoa.org)
Git hash: 1b39c1bacfc1b07f96d689db230b2586549a62d4