

Optimal decision and naive Bayes

Fabrice Rossi

SAMM
Université Paris 1 Panthéon Sorbonne

2017

Standard supervised learning hypothesis

Optimal model

The Naive Bayes Classifier

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”:

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

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: $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{iP})^T$

Risk of a model

The risk of g for the loss function l is

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

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
- ▶ stationarity also for e.g. drift analysis

Standard supervised learning hypothesis

Optimal model

The Naive Bayes Classifier

Optimal risk

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

Is R_I^* reachable?

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

Optimal risk

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

Is R_I^* reachable?

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

Quadratic case

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

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

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

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

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

- ▶ standard properties of the expectation

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

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

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

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

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

Ratios of probabilities are sufficient to compute the optimal model

Standard supervised learning hypothesis

Optimal model

The Naive Bayes Classifier

Generative model

- ▶ a model that explains both \mathbf{X} and \mathbf{Y}
- ▶ as opposed to \mathbf{Y} given \mathbf{X}
- ▶ one road to optimal models

Hypotheses

- ▶ $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \dots \times \mathcal{X}_p$
- ▶ \mathcal{Y} and all the \mathcal{X}_j are finite sets

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

- ▶ $|\mathcal{X}_1| \times |\mathcal{X}_2| \times \dots \times |\mathcal{X}_p| \times |\mathcal{Y}| - 1$ parameters
- ▶ generally intractable

Conditional independence

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

Categorical distribution

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

Naive Bayes distribution

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

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

Maximum Likelihood Estimation

MLE

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

Naive Bayes log Likelihood

- ▶ standard i.i.d. assumptions
- ▶ separability

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

Separate estimations

- ▶ parameters of the different categorical distributions are independent
- ▶ $\widehat{\Gamma}_{MLE}$ 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

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)

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 $NB(\Gamma)$ for a data set
- ▶ approximate the data distribution by the NB distribution i.e.

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

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

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

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

Naive Bayes Classifier

Example

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

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

and thus

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

- ▶ very simple frequency comparison!

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