Efficient Learning of Naive Bayes Classifiers under Class-Conditional Classification Noise

François Denis, Christophe Magnan, Liva Ralaivola

Laboratoire d'Informatique Fondamentale de Marseille (LIF)

ICML 2006

Outline

Statistical Learning Framework

$X = \prod_{i=1}^{m} X^i$, a domain defined by m symbolic attributes
$Y = \{0, 1\}$, classes

Data: $S = \{(x_1, y_1), ..., (x_l, y_l)\}$ i.i.d. wrt $P(x, y) = P(x) \cdot P(y|x)$

Statistical Learning Framework

$X = \prod_{i=1}^{m} X^i$, a domain defined by m symbolic attributes
$Y = \{0, 1\}$, classes

Data: $S = \{(x_1, y_1), \dots, (x_l, y_l)\}$ i.i.d. wrt $P(x, y) = P(x) \cdot P(y|x)$

Goal: compute a classifier $f : X \rightarrow Y$ with low risk $R(f) = P(f(x) \neq y)$.

Statistical Learning Framework

$X = \prod_{i=1}^{m} X^i$, a domain defined by m symbolic attributes
$Y = \{0, 1\}$, classes

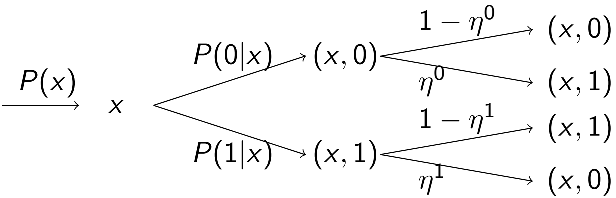
Data: $S = \{(x_1, y_1), \ldots, (x_l, y_l)\}$ i.i.d. wrt $P(x, y) = P(x) \cdot P(y|x)$

Goal: compute a classifier $f : X \to Y$ with low risk $R(f) = P(f(x) \neq y)$.

Bayes classifier: $f_{Bayes}(x) = argmax_y P(y|x)$

Class Conditional Classification Noise (CCCN)

Let $\overrightarrow{\eta} = [\eta^0 \ \eta^1]$ where $\eta^0, \eta^1 \in [0, 1]$



Additional noise rates only depend on class labels.

$$\begin{cases} P^{\overrightarrow{\eta}}(0|x) = (1-\eta^0) \cdot P(0|x) + \eta^1 \cdot P(1|x) \\ P^{\overrightarrow{\eta}}(1|x) = (1-\eta^1) \cdot P(1|x) + \eta^0 \cdot P(0|x) \end{cases}$$

$P^{\overrightarrow{\eta}}(x, y) = P(x) P^{\overrightarrow{\eta}}(y|x)$: the noisy joint distribution.

Remark

$$\begin{cases} P^{\overrightarrow{\eta}}(0|x) = (1-\eta^0)\cdot P(0|x) + \eta^1 \cdot P(1|x) \\ P^{\overrightarrow{\eta}}(1|x) = (1-\eta^1)\cdot P(1|x) + \eta^0 \cdot P(0|x) \\ P^{\overrightarrow{\eta}}(x,y) = P(x)P^{\overrightarrow{\eta}}(y|x) \end{cases}$$

Remark

$$\begin{cases} P^{\vec{\eta}}(0|x) = (1-\eta^0)\cdot P(0|x) + \eta^1 \cdot P(1|x) \\ P^{\vec{\eta}}(1|x) = (1-\eta^1)\cdot P(1|x) + \eta^0 \cdot P(0|x) \\ P^{\vec{\eta}}(x,y) = P(x)P^{\vec{\eta}}(y|x) \end{cases}$$

- $\eta^0 + \eta^1 = 1 \Rightarrow P^{\vec{\eta}}(0|x) = \eta^1 \wedge P^{\vec{\eta}}(1|x) = \eta^0$.

Nothing can be learned from $P^{\vec{\eta}}$.

Remark

$$\begin{cases} P^{\vec{\eta}}(0|x) = (1-\eta^0)\cdot P(0|x) + \eta^1 \cdot P(1|x) \\ P^{\vec{\eta}}(1|x) = (1-\eta^1)\cdot P(1|x) + \eta^0 \cdot P(0|x) \\ P^{\vec{\eta}}(x,y) = P(x)P^{\vec{\eta}}(y|x) \end{cases}$$

- $\eta^0 + \eta^1 = 1 \Rightarrow P^{\vec{\eta}}(0|x) = \eta^1 \wedge P^{\vec{\eta}}(1|x) = \eta^0$.

  Nothing can be learned from $P^{\vec{\eta}}$.

- $P'(x,y) = P(x, 1-y), \eta'^0 = 1-\eta^1, \eta'^1 = 1-\eta^0$.

  $P'^{\vec{\eta}'} = P^{\vec{\eta}}$ while $f'_{Bayes} = 1 - f_{Bayes}$.

  P and P' cannot be distinguished.

Remark

$$\begin{cases} P^{\vec{\eta}}(0|x) = (1-\eta^0)\cdot P(0|x) + \eta^1 \cdot P(1|x) \\ P^{\vec{\eta}}(1|x) = (1-\eta^1)\cdot P(1|x) + \eta^0 \cdot P(0|x) \\ P^{\vec{\eta}}(x,y) = P(x)P^{\vec{\eta}}(y|x) \end{cases}$$

- $\eta^0 + \eta^1 = 1 \Rightarrow P^{\vec{\eta}}(0|x) = \eta^1 \wedge P^{\vec{\eta}}(1|x) = \eta^0$.

  Nothing can be learned from $P^{\vec{\eta}}$.

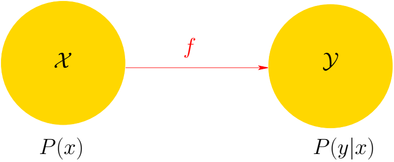
- $P'(x,y) = P(x, 1-y), \eta'^0 = 1-\eta^1, \eta'^1 = 1-\eta^0$.

  $$P'^{\vec{\eta}'} = P^{\vec{\eta}} \text{ while } f'_{Bayes} = 1 - f_{Bayes}.$$

  P and P' cannot be distinguished.

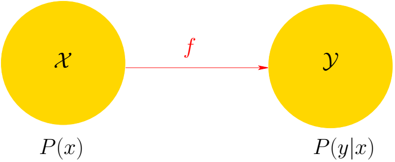
From now on, we suppose that $\eta^0 + \eta^1 < 1$.

Learning under Class Conditional Classification Noise



Data: $S^{\overrightarrow{\eta}} = \{(x_1, y_1), ..., (x_l, y_l)\}$ i.i.d. wrt
$P^{\overrightarrow{\eta}}(x, y) = P(x) \cdot P^{\overrightarrow{\eta}}(y|x)$

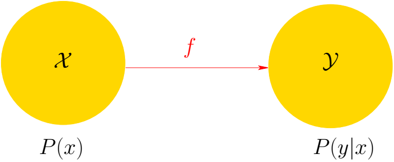
Learning under Class Conditional Classification Noise



Data: $S^{\overrightarrow{\eta}} = \{(x_1, y_1), \ldots, (x_l, y_l)\}$ i.i.d. wrt $P^{\overrightarrow{\eta}}(x, y) = P(x) \cdot P^{\overrightarrow{\eta}}(y|x)$

Goal: compute a classifier $f : X \to Y$ which minimizes $R(f) = P(f(x) \neq y)$ (not $R^{\overrightarrow{\eta}}(f) = P^{\overrightarrow{\eta}}(f(x) \neq y)$!)

Learning under Class Conditional Classification Noise



Data: $S^{\overrightarrow{\eta}} = \{(x_1, y_1), \ldots, (x_l, y_l)\}$ i.i.d. wrt $P^{\overrightarrow{\eta}}(x, y) = P(x) \cdot P^{\overrightarrow{\eta}}(y|x)$

Goal: compute a classifier $f : X \rightarrow Y$ which minimizes $R(f) = P(f(x) \neq y)$ (not $R^{\overrightarrow{\eta}}(f) = P^{\overrightarrow{\eta}}(f(x) \neq y)$!)

Is it possible to learn under CCCN as well as without noise?

P and $P^{\vec{\eta}}$ can define the same Bayes classifier

$$P(1|x) \geq P(0|x) \Leftrightarrow P^{\vec{\eta}}(1|x) \geq P^{\vec{\eta}}(0|x)$$

iff

$$P(1|x) \geq P(0|x) \Leftrightarrow (1-2\eta^1)\cdot P(1|x) \geq (1-2\eta^0)\cdot P(0|x)$$

P and $P^{\vec{\eta}}$ can define the same Bayes classifier

$$P(1|x) \geq P(0|x) \Leftrightarrow P^{\vec{\eta}}(1|x) \geq P^{\vec{\eta}}(0|x)$$

iff

$$P(1|x) \geq P(0|x) \Leftrightarrow (1 - 2\eta^1) \cdot P(1|x) \geq (1 - 2\eta^0) \cdot P(0|x)$$

- Uniform classification noise:

$$\eta^0 = \eta^1 < 1/2 \Rightarrow f_{Bayes} = f^{\vec{\eta}}_{Bayes}$$

P and $P^{\vec{\eta}}$ can define the same Bayes classifier

$$P(1|x) \geq P(0|x) \Leftrightarrow P^{\vec{\eta}}(1|x) \geq P^{\vec{\eta}}(0|x)$$

iff

$$P(1|x) \geq P(0|x) \Leftrightarrow (1 - 2\eta^1) \cdot P(1|x) \geq (1 - 2\eta^0) \cdot P(0|x)$$

- Uniform classification noise:

$$\eta^0 = \eta^1 < 1/2 \Rightarrow f_{Bayes} = f^{\vec{\eta}}_{Bayes}$$

- Deterministic target:

$$[\forall x, (P(1|x) = 0 \text{ or } P(0|x) = 0) \text{ and } \eta^0, \eta^1 < 1/2] \Rightarrow f_{Bayes} = f^{\vec{\eta}}_{Bayes}$$

General case: an ill-posed problem?

Let $X = \{a\}$, P_1 and P_2 such that

- $P_1(0|a) = \frac{1}{3} \Rightarrow f_{Bayes}(a) = 1$
- $\overrightarrow{\eta}_1 = (0,0) \Rightarrow P_1^{\overrightarrow{\eta}}(0|a) = \frac{1}{3}$

- $P_2(0|a) = \frac{2}{3} \Rightarrow f_{Bayes}(a) = 0$
- $\overrightarrow{\eta}_2 = (\frac{1}{2}, 0) \Rightarrow P_2^{\overrightarrow{\eta}}(0|a) = \frac{1}{3}$

- $P_1^{\overrightarrow{\eta}_1} = P_2^{\overrightarrow{\eta}_2}$
- $f_{1,Bayes} = 1 - f_{2,Bayes}$

Identifiability under CCCN

$\mathcal{P}$: a set of probability distributions over $X \times Y$.

$\mathcal{P}$ is *identifiable under CCCN*
if
$\forall P \in \mathcal{P}, \forall \eta^0, \eta^1$ s.t. $\eta^0 + \eta^1 < 1$, $P^{\overrightarrow{\eta}}$ determines P.

$$P_1^{\overrightarrow{\eta}_1} = P_2^{\overrightarrow{\eta}_2} \Rightarrow P_1 = P_2 \text{ and } \overrightarrow{\eta}_1 = \overrightarrow{\eta}_2.$$

Identifiability under CCCN: a simple case

$\mathcal{Q}$: a set of distributions over X

Def. The 2-mixtures of elements of $\mathcal{Q}$ are *identifiable* if
$\forall Q_1, Q_2 \in \mathcal{Q}, \alpha \in [0,1]$,
$\alpha Q_1 + (1-\alpha)Q_2$ determines α, Q_1 and Q_2 (up to a permutation).

Theorem. Let $\mathcal{P}$ be a set of distributions over $X \times Y$,
let $\mathcal{Q} = \{P(\cdot|y) | y \in Y, P \in \mathcal{P}\}$.
If the 2-mixtures of $\mathcal{Q}$ are identifiable, then $\mathcal{P}$ is *identifiable under CCCN*.

Proof. Let $P \in \mathcal{P}$ and η^0, η^1.

- $P^{\vec{\eta}}(x|1) = \alpha P(x|1) + (1-\alpha)P(x|0)$
- $P^{\vec{\eta}}(x|0) = \beta P(x|1) + (1-\beta)P(x|0)$
- $P(1) = \beta + (\alpha - \beta)P^{\vec{\eta}}(1)$
- $P(x,1) = P(x|1)P(1)$ and $P(x,0) = P(x|0)P(0)$.

Outline

Product distributions

Let $X = \prod_{i=1}^{m} X^i$ and let Q be a probability distribution over X.

Q is a *product distribution* if $Q(x) = \prod_{i=1}^{m} Q(x^i)$.

Theorem. Mixtures of product distributions are identifiable [GHKM01, WT02, FM99, FDS05].

Def. *Naive Bayes distributions:* P over $X \times Y$ such that $P(\cdot|0)$ and $P(\cdot|1)$ are product distributions.

Cor. The set of naive Bayes distributions over $X \times Y$ is identifiable under CCCN.

2-mixtures of 2 product distributions: an analytical identification.

Let $\begin{cases} Q_\alpha = \alpha P_1 + (1-\alpha)P_2 \\ Q_\beta = \beta P_1 + (1-\beta)P_2 \end{cases}$ where $\alpha \neq \beta$.

$$C = (Q_\alpha(x^i = a) - Q_\beta(x^i = a))(Q_\alpha(x^j = b) - Q_\beta(x^j = b))$$
$$D = Q_\alpha(a, b) - Q_\alpha(x^i = a)Q_\alpha(x^j = b)$$

General case: $\beta \neq 0$ **and** $\beta \neq 1$

$$E = Q_\beta(a, b) - Q_\beta(x^i = a)Q_\beta(x^j = b)$$
$$\lambda_\beta = \frac{CE}{(C+D+E)^2 - 4DE}$$
$$\beta^2 - \beta + \lambda_\beta = 0 \text{ and } \alpha = \beta \cdot \frac{(1-\beta)(C+D) - \beta E}{E(1-2\beta)}$$

2-mixtures of 2 product distributions: an analytical identification.

Let $\begin{cases} Q_\alpha = \alpha P_1 + (1-\alpha)P_2 \\ Q_\beta = \beta P_1 + (1-\beta)P_2 \end{cases}$ where $\alpha \neq \beta$.

$$C = (Q_\alpha(x^i = a) - Q_\beta(x^i = a))(Q_\alpha(x^j = b) - Q_\beta(x^j = b))$$
$$D = Q_\alpha(a, b) - Q_\alpha(x^i = a)Q_\alpha(x^j = b)$$

Particular case: $\beta = 0$ **or** $\beta = 1$

- If $\beta = 0$ then $\alpha = \frac{C}{D+C}$,
- If $\beta = 1$ then $\alpha = \frac{D}{D+C}$

2-mixtures of 2 product distributions: an analytical identification.

Let $\begin{cases} Q_\alpha = \alpha P_1 + (1-\alpha)P_2 \\ Q_\beta = \beta P_1 + (1-\beta)P_2 \end{cases}$ where $\alpha \neq \beta$.

$$C = (Q_\alpha(x^i = a) - Q_\beta(x^i = a))(Q_\alpha(x^j = b) - Q_\beta(x^j = b))$$
$$D = Q_\alpha(a, b) - Q_\alpha(x^i = a)Q_\alpha(x^j = b)$$

In all cases:

$$P_1 = \frac{(1-\beta)Q_\alpha - (1-\alpha)Q_\beta}{\alpha - \beta} \text{ and } P_2 = \frac{-\beta Q_\alpha + \alpha Q_\beta}{\alpha - \beta}$$

Application to Naive Bayes distributions.

Let P be a naive Bayes distribution over $X \times Y$.
$P^{\vec{\eta}}(x|1) = \alpha P(x|1) + (1-\alpha)P(x|0)$ and
$P^{\vec{\eta}}(x|0) = \beta P(x|1) + (1-\beta)P(x|0)$.

$C = (P^{\vec{\eta}}(x^i = a|1) - P^{\vec{\eta}}(x^i = a|0))(P^{\vec{\eta}}(x^j = b|1) - P^{\vec{\eta}}(x^j = b|0))$
$D = P^{\vec{\eta}}(x^i = a, x^j = b|1) - P^{\vec{\eta}}(x^i = a|1)P^{\vec{\eta}}(x^j = b|1)$
$E = P^{\vec{\eta}}(x^i = a, x^j = b|0) - P^{\vec{\eta}}(x^i = a|0)P^{\vec{\eta}}(x^j = b|0)$

$\lambda_\beta = \frac{CE}{(C+D+E)^2 - 4DE}$.

$$\beta^2 - \beta + \lambda_\beta = 0 \text{ and } \alpha = \beta \cdot \frac{(1-\beta)(C+D) - \beta E}{E(1-2\beta)}$$

$P(1) = \beta + (\alpha - \beta)P^{\vec{\eta}}(1)$ and $P(x) = P(1)P(x|1) + (1-P(1))P(x|0)$.

Learning Naive Bayes distributions.

Let P be a naive Bayes distribution over $X \times Y$.
$P^{\overrightarrow{\eta}}(x|1) = \alpha P(x|1) + (1-\alpha)P(x|0)$ and
$P^{\overrightarrow{\eta}}(x|0) = \beta P(x|1) + (1-\beta)P(x|0)$.

$$\hat{C}_{i,j}^{a,b} = (\widehat{P^{\overrightarrow{\eta}}}(x_i = a|1) - \widehat{P^{\overrightarrow{\eta}}}(x_i = a|0))(\widehat{P^{\overrightarrow{\eta}}}(x_j = b|1) - \widehat{P^{\overrightarrow{\eta}}}(x_j = b|0))$$
$$\hat{D}_{i,j}^{a,b} = \widehat{P^{\overrightarrow{\eta}}}(x_i = a, x_j = b|1) - \widehat{P^{\overrightarrow{\eta}}}(x_i = a|1)\widehat{P^{\overrightarrow{\eta}}}(x_j = b|1)$$
$$\hat{E}_{i,j}^{a,b} = \widehat{P^{\overrightarrow{\eta}}}(x_i = a, x_j = b|0) - \widehat{P^{\overrightarrow{\eta}}}(x_i = a|0)\widehat{P^{\overrightarrow{\eta}}}(x_j = b|0)$$

$$\hat{\lambda}_\beta = \frac{\sum \hat{C}_{i,j}^{a,b}\hat{E}_{i,j}^{a,b}}{\sum[(\hat{C}_{i,j}^{a,b}+\hat{D}_{i,j}^{a,b}+\hat{E}_{i,j}^{a,b})^2 - 4\hat{D}_{i,j}^{a,b}\hat{E}_{i,j}^{a,b}]} \text{ and } \hat{\lambda}_\alpha = \frac{\sum \hat{C}_{i,j}^{a,b}\hat{D}_{i,j}^{a,b}}{\sum[(\hat{C}_{i,j}^{a,b}+\hat{D}_{i,j}^{a,b}+\hat{E}_{i,j}^{a,b})^2 - 4\hat{D}_{i,j}^{a,b}\hat{E}_{i,j}^{a,b}]}$$

$$\hat{\beta}^2 - \hat{\beta} + \hat{\lambda}_\beta = 0 \text{ and } \hat{\alpha}^2 - \hat{\alpha} + \hat{\lambda}_\alpha = 0$$

$$\hat{P}(1) = \hat{\beta} + (\hat{\alpha} - \hat{\beta})\widehat{P^{\overrightarrow{\eta}}}(1).$$

Application to Semisupervised Learning from Positive and Unlabeled Examples.

- $P(x)$: unlabeled examples
- $P(x|1)$: positive examples.

A crucial parameter: $P(1)$

- $P(x,1) = P(x|1)P(1)$
- $P(x,0) = P(x) - P(x|1)P(1)$.

But in general, $P(1)$ has to be provided to the algorithm.

Application to Semisupervised Learning from Positive and Unlabeled Examples.

Let P be a naive Bayes distribution over $X \times Y$.

$P(x) = \alpha P(x|1) + (1-\alpha)P(x|0)$ where $\alpha = P(1)$.
$P(x|1) = \beta P(x|1) + (1-\beta)P(x|0)$ where $\beta = 1$.

$C = (P(x^i = a) - P(x^i = a|1))(P(x^j = b) - P(x^j = b|1))$
$D = P(x^i = a, x^j = b) - P(x^i = a)P(x^j = b)$

$$\alpha = P(1) = \frac{D}{C + D}$$

Application to Semisupervised Learning from Positive and Unlabeled Examples.

Let P be a naive Bayes distribution over $X \times Y$.

- $P(1)$ is determined by $P(x)$ and $P(x|1)$
- $P(1)$ can be estimated from samples of positive and unlabeled examples.

Outline

Algorithms

- `NB-CCCN`: directly deduced from the analytical formulas.

Algorithms

- `NB-CCCN`: directly deduced from the analytical formulas.
- `NB-CCCN-EM`: starting from the model θ_0 output by `NB-CCCN`, maximizing the likelihood of the learning data.
  - Given a model θ_k, estimate the probability that the label of a given example has been corrupted,
  - Use these estimates to compute a model θ_{k+1}.

Algorithms

- `NB-CCCN`: directly deduced from the analytical formulas.
- `NB-CCCN-EM`: starting from the model θ_0 output by `NB-CCCN`, maximizing the likelihood of the learning data.
  - Given a model θ_k, estimate the probability that the label of a given example has been corrupted,
  - Use these estimates to compute a model θ_{k+1}.
- `NB-UNL`: from unlabeled data, using the analytical formulas provided in [GHKM01].

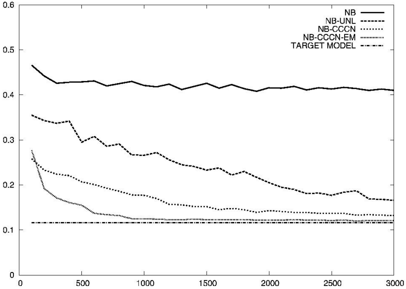
Algorithms

- `NB-CCCN`: directly deduced from the analytical formulas.
- `NB-CCCN-EM`: starting from the model θ_0 output by `NB-CCCN`, maximizing the likelihood of the learning data.
  - Given a model θ_k, estimate the probability that the label of a given example has been corrupted,
  - Use these estimates to compute a model θ_{k+1}.
- `NB-UNL`: from unlabeled data, using the analytical formulas provided in [GHKM01].
- `NB`: standard naive Bayes algorithm.

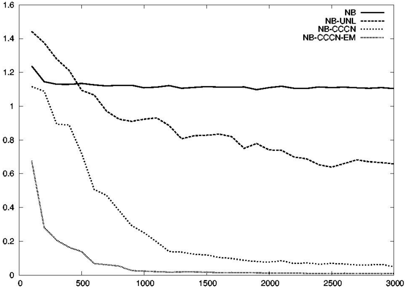
Artificial data

- 10 binary attributes,
- targets: randomly drawn naive Bayes distributions,
- noise rates: $\eta^0 = 0.2$ et $\eta^1 = 0.5$,
- average of 200 experiments,
- test sets: 10,000 examples **not corrupted by noise**.

Results on artificial data



$R_{emp}^{test}(\theta)$



$d_{KL}(\theta, \theta_t)$

Experiments on UCI data sets

Nom	$\lvert S \rvert$	$NbAtt$	$\lvert X^i \rvert$
House Votes	433	16	2
Tic Tac Toe	958	9	3
Hepatitis	155	19	2-10
Breast Cancer	286	9	2-11
B. C. Wisc.	699	9	10
Bal. Scale	576	4	5

10-folds cross Validation
Noise added to the learning set: [0,0] and [0.2, 0.5].
No noise on test sets.

Results on UCI data sets

Dataset		MC	NB	NB-CCCN	NB-CCCN-EM
H.Votes	ac	0.62	0.904	**0.916**	0.882
no noise	lk	-	-3134	-3035	**-2915**
	$\overrightarrow{\hat{\eta}}$	-	-	(.02,.08)	(.04,.20)
H.Votes	ac	0.38	0.866	**0.900**	0.873
$\overrightarrow{\eta}$ noise	lk	-	-4130	**-3037**	-3041
	$\overrightarrow{\hat{\eta}}$	-	-	(.33,.58)	(.20,.56)
T.T.T.	ac	0.65	**0.697**	0.682	**0.697**
no noise	lk	-	**-8726**	-8854	**-8726**
	$\overrightarrow{\hat{\eta}}$	-	-	(.09,.19)	(.00,.00)
T.T.T.	ac	0.35	0.562	**0.664**	0.587
$\overrightarrow{\eta}$ noise	lk	-	-8828	-8818	**-8815**
	$\overrightarrow{\hat{\eta}}$	-	-	(.24,.62)	(.21,.56)

Results on UCI data sets

Dataset		MC	NB	NB-CCCN	NB-CCCN-EM
Hepat. no noise	ac	0.79	0.827	**0.850**	0.770
	lk	-	-1982	-2416	**-1902**
	$\overrightarrow{\hat{\eta}}$	-	-	(.31,.03)	(.50,.03)
Hepat. $\overrightarrow{\eta}$ noise	ac	0.21	0.590	**0.811**	0.758
	lk	-	-2095	-2273	**-1946**
	$\overrightarrow{\hat{\eta}}$	-	-	(.25,.55)	(.29,.45)
Br.Can. no noise	ac	0.70	0.730	**0.760**	0.718
	lk	-	-2520	-2682	**-2448**
	$\overrightarrow{\hat{\eta}}$	-	-	(.06,.20)	(.13,.27)
Br.Can. $\overrightarrow{\eta}$ noise	ac	0.30	0.581	**0.732**	0.722
	lk	-	-2573	-2623	**-2479**
	$\overrightarrow{\hat{\eta}}$	-	-	(.19,.59)	(.33,.56)

Outline

1. Learning under CCC-noise
2. Learning Naive Bayes classifiers under CCCN
3. Experiments
4. Conclusion

Conclusions and prospects

- CN learnability in the statistical learning framework
- Naive Bayes distributions are identifiable under CCCN
- Which classes of distributions are identifiable under CCCN?
- Naive Bayes distributions are learnable under CCCN
- Convergence rates?
- Minimizing the empirical risk on noisy data is not (always) a consistent strategy.

$$R(f) = \frac{R^{\overrightarrow{\eta}}(f) - \eta^1 \cdot p_f - \eta^0 \cdot (1 - p_f)}{1 - \eta^0 - \eta^1}$$

On which empirical measure can we rely?