VAPNIK-CHERVONENKIS THEORY IN PATTERN RECOGNITION

András Antos

BMGE, MIT, Intelligent Data Analysis, Apr 12, 2018
Based on: [Devroye et al., 1996], PDSS, IDA jegyzet

OUTLINE

1. INTRO: DECISION, SUPERVISED (PASSIVE) LEARNING
   - Bayes decision
   - Approximation of Bayes decision
   - Sample based classification
   - No rate - Slow rate of convergence
   - Resticted class - Empirical risk minimization

OUTLINE

1. INTRO: DECISION, SUPERVISED (PASSIVE) LEARNING
   - Bayes decision
   - Approximation of Bayes decision
   - Sample based classification
   - No rate - Slow rate of convergence
   - Resticted class - Empirical risk minimization

DECISION PROBLEM, ERROR PROBABILITY

- Decide for not (yet) observable Y based on an observable X
- X, Y r.v.'s, with domains $\mathcal{X}$ (e.g. $\subseteq \mathbb{R}^d$) and $\mathcal{Y} = \{0, 1\}$ *labels*, resp., and with joint distr. ν
- $g : \mathcal{X} \to \mathcal{Y}$ *decision function* or *classifier* is used to decide from X to Y
- Goodness of $g(X)$ *decision* is measured by *0-1* cost: 1, if $g(X)$ differs from true Y, else 0 $\Rightarrow$
- Performance of g is measured by its error probability (global risk): $R(g) \stackrel{\text{def}}{=} \mathbb{P}(Y \neq g(X))$
- g's minimizing $R(g)$: *optimal*

DECISION PROBLEM, ERROR PROBABILITY

- Decide for not (yet) observable Y based on an observable X
- X, Y r.v.'s, with domains $\mathcal{X}$ (e.g. $\subseteq \mathbb{R}^d$) and $\mathcal{Y} = \{0, 1\}$ *labels*, resp., and with joint distr. ν
- $g : \mathcal{X} \to \mathcal{Y}$ *decision function* or *classifier* is used to decide from X to Y
- Goodness of $g(X)$ *decision* is measured by *0-1* cost: 1, if $g(X)$ differs from true Y, else 0 $\Rightarrow$
- Performance of g is measured by its error probability (global risk): $R(g) \stackrel{\text{def}}{=} \mathbb{P}(Y \neq g(X))$
- g's minimizing $R(g)$: *optimal*

DECISION PROBLEM, ERROR PROBABILITY

- Decide for not (yet) observable Y based on an observable X
- X, Y r.v.'s, with domains $\mathcal{X}$ (e.g. $\subseteq \mathbb{R}^d$) and $\mathcal{Y} = \{0, 1\}$ *labels*, resp., and with joint distr. ν
- $g : \mathcal{X} \to \mathcal{Y}$ *decision function* or *classifier* is used to decide from X to Y
- Goodness of $g(X)$ *decision* is measured by *0-1* cost: 1, if $g(X)$ differs from true Y, else 0 $\Rightarrow$
- Performance of g is measured by its error probability (global risk): $R(g) \stackrel{\text{def}}{=} \mathbb{P}(Y \neq g(X))$
- g's minimizing $R(g)$: *optimal*

DECISION PROBLEM, ERROR PROBABILITY

- Decide for not (yet) observable Y based on an observable X
- X, Y r.v.'s, with domains $\mathcal{X}$ (e.g. $\subseteq \mathbb{R}^d$) and $\mathcal{Y} = \{0, 1\}$ *labels*, resp., and with joint distr. ν
- $g : \mathcal{X} \to \mathcal{Y}$ *decision function* or *classifier* is used to decide from X to Y
- Goodness of $g(X)$ *decision* is measured by *0-1* cost: 1, if $g(X)$ differs from true Y, else 0 $\Rightarrow$
- Performance of g is measured by its error probability (global risk): $R(g) \stackrel{\text{def}}{=} \mathbb{P}(Y \neq g(X))$
- g's minimizing $R(g)$: *optimal*

DECISION PROBLEM, ERROR PROBABILITY

- Decide for not (yet) observable Y based on an observable X
- X, Y r.v.'s, with domains $\mathcal{X}$ (e.g. $\subseteq \mathbb{R}^d$) and $\mathcal{Y} = \{0, 1\}$ *labels*, resp., and with joint distr. ν
- $g : \mathcal{X} \to \mathcal{Y}$ *decision function* or *classifier* is used to decide from X to Y
- Goodness of $g(X)$ *decision* is measured by *0-1* cost: 1, if $g(X)$ differs from true Y, else 0 $\Rightarrow$
- Performance of g is measured by its error probability (global risk): $R(g) \stackrel{\text{def}}{=} \mathbb{P}(Y \neq g(X))$
- g's minimizing $R(g)$: *optimal*

DECISION PROBLEM, ERROR PROBABILITY

- Decide for not (yet) observable Y based on an observable X
- X, Y r.v.'s, with domains $\mathcal{X}$ (e.g. $\subseteq \mathbb{R}^d$) and $\mathcal{Y} = \{0, 1\}$ *labels*, resp., and with joint distr. ν
- $g : \mathcal{X} \to \mathcal{Y}$ *decision function* or *classifier* is used to decide from X to Y
- Goodness of $g(X)$ *decision* is measured by *0-1* cost: 1, if $g(X)$ differs from true Y, else 0 $\Rightarrow$
- Performance of g is measured by its error probability (global risk): $R(g) \stackrel{\text{def}}{=} \mathbb{P}\,(Y \neq g(X))$
- g's minimizing $R(g)$: *optimal*

HYPOTHESIS, DECISION DOMAIN

DEFINITIONS

For $i = 0{,}1$: $\{Y = i\}$: i^{th} *hypothesis*. Y a posteriori distribution is given by $\eta_i(x) \stackrel{\text{def}}{=} \mathbb{P}(Y = i|X = x)$ *a posteriori probabilities*. Preimages of 0 and 1 by g form a partition of $\mathcal{X}$, its classes $D_i = \{x \in \mathcal{X} : g(x) = i\}$ are the *decision domains*.

Note: $1 - \eta_0(x) = \eta_1(x) = \mathbb{E}[Y|X = x] \stackrel{\text{def}}{=} \eta(x)$ (*regression* or *a posteriori probability* function).
If $X \sim \mu$, ν may be given, e.g., by (μ, η). $\forall C_0, C_1 \subseteq \mathcal{X}$

$$\mathbb{P}((X, Y) \in C_0 \times \{0\} \cup C_1 \times \{1\}) = \int_{C_0} (1 - \eta) d\mu + \int_{C_1} \eta \, d\mu.$$

$(D_0, D_1) \Leftrightarrow g$, since $\mathbb{I}_{\{g(x)=j\}} = \mathbb{I}_{\{x \in D_j\}}$ ($\mathbb{I}_A$: indicator func. of A).

LOCAL RISK

$r(g,x) \stackrel{\text{def}}{=} \mathbb{P}\left(Y \neq g(X) | X = x\right)$: *local risk function*
$\mathbb{E}\left[r(g,X)\right] = R(g)$ and

$$\begin{aligned} r(g,x) &= \mathbb{I}_{\{g(x)=1\}}\eta_0(x) + \mathbb{I}_{\{g(x)=0\}}\eta_1(x) \\ &= 1 - \mathbb{I}_{\{x\in D_0\}}\eta_0(x) - \mathbb{I}_{\{x\in D_1\}}\eta_1(x). \end{aligned}$$

Minimized by g which puts $\forall x$ into D_j with the greater $\eta_j(x)$.

OUTLINE

BAYES DECISION

- Let $\{D_j^*\}$ be s.t. $\forall x$

$$x \in D_j^* \Leftrightarrow (\eta_j(x) > \eta_{1-j}(x) \text{ v. } j = 0, \eta_0(x) = \eta_1(x))$$

g^* picks the more likely j given X.

$$x \in D_j^* \Rightarrow \eta_j(x) = \max(\eta_0(x), \eta_1(x)).$$

DEFINITION

Bayes decision (*maximum a posteriori decision*): g^* corresp. to (D_0^*, D_1^*) above, i.e. $g^*(x) = 1 \Leftrightarrow x \in D_1^* \Leftrightarrow \eta(x) > 1/2$.

THEOREM

The Bayes decision minimizes $r(g, x)$ $\forall x$, and so optimal. The minimum is $r(g^*, x) = \min(\eta_0(x), \eta_1(x))$.

(optimal) global risk of g^*: *Bayes risk/Bayes error*

$$R^* \stackrel{\text{def}}{=} R(g^*) = \mathbb{E}[\min(\eta_0(X), \eta_1(X))] = \mathbb{E}[\min(\eta(X), 1 - \eta(X))].$$

BAYES DECISION

- Let $\{D_j^*\}$ be s.t. $\forall x$

$$x \in D_j^* \Leftrightarrow (\eta_j(x) > \eta_{1-j}(x) \text{ v. } j = 0, \eta_0(x) = \eta_1(x))$$

g^* picks the more likely j given X.

$$x \in D_j^* \Rightarrow \eta_j(x) = \max(\eta_0(x), \eta_1(x)).$$

DEFINITION

Bayes decision (*maximum a posteriori decision*): g^* corresp. to (D_0^*, D_1^*) above, i.e. $g^*(x) = 1 \Leftrightarrow x \in D_1^* \Leftrightarrow \eta(x) > 1/2$.

THEOREM

The Bayes decision minimizes $r(g, x)$ $\forall x$, and so optimal. The minimum is $r(g^*, x) = \min(\eta_0(x), \eta_1(x))$.

(optimal) global risk of g^*: *Bayes risk/Bayes error*

$$R^* \stackrel{\text{def}}{=} R(g^*) = \mathbb{E}[\min(\eta_0(X), \eta_1(X))] = \mathbb{E}[\min(\eta(X), 1 - \eta(X))].$$

BAYES DECISION

- Let $\{D_j^*\}$ be s.t. $\forall x$

$$x \in D_j^* \Leftrightarrow (\eta_j(x) > \eta_{1-j}(x) \text{ v. } j = 0, \eta_0(x) = \eta_1(x))$$

g^* picks the more likely j given X.

$$x \in D_j^* \Rightarrow \eta_j(x) = \max(\eta_0(x), \eta_1(x)).$$

DEFINITION

Bayes decision (*maximum a posteriori decision*): g^* corresp. to (D_0^*, D_1^*) above, i.e. $g^*(x) = 1 \Leftrightarrow x \in D_1^* \Leftrightarrow \eta(x) > 1/2$.

THEOREM

The Bayes decision minimizes $r(g, x)$ $\forall x$, and so optimal. The minimum is $r(g^*, x) = \min(\eta_0(x), \eta_1(x))$.

(optimal) global risk of g^*: *Bayes risk*/*Bayes error*

$$R^* \stackrel{\text{def}}{=} R(g^*) = \mathbb{E}[\min(\eta_0(X), \eta_1(X))] = \mathbb{E}[\min(\eta(X), 1 - \eta(X))].$$

BAYES DECISION

- Let $\{D_j^*\}$ be s.t. $\forall x$

$$x \in D_j^* \Leftrightarrow (\eta_j(x) > \eta_{1-j}(x) \text{ v. } j = 0, \eta_0(x) = \eta_1(x))$$

g^* picks the more likely j given X.

$$x \in D_j^* \Rightarrow \eta_j(x) = \max(\eta_0(x), \eta_1(x)).$$

DEFINITION

Bayes decision (*maximum a posteriori decision*): g^* corresp. to (D_0^*, D_1^*) above, i.e. $g^*(x) = 1 \Leftrightarrow x \in D_1^* \Leftrightarrow \eta(x) > 1/2$.

THEOREM

The Bayes decision minimizes $r(g, x)$ $\forall x$, and so optimal. The minimum is $r(g^*, x) = \min(\eta_0(x), \eta_1(x))$.

(optimal) global risk of g^*: *Bayes risk*/*Bayes error*

$$R^* \stackrel{\text{def}}{=} R(g^*) = \mathbb{E}\left[\min(\eta_0(X), \eta_1(X))\right] = \mathbb{E}\left[\min(\eta(X), 1 - \eta(X))\right].$$

BAYES DECISION

- Let $\{D_j^*\}$ be s.t. $\forall x$

$$x \in D_j^* \Leftrightarrow (\eta_j(x) > \eta_{1-j}(x) \text{ v. } j = 0, \eta_0(x) = \eta_1(x))$$

g^* picks the more likely j given X.

$$x \in D_j^* \Rightarrow \eta_j(x) = \max(\eta_0(x), \eta_1(x)).$$

DEFINITION

Bayes decision (*maximum a posteriori decision*): g^* corresp. to (D_0^*, D_1^*) above, i.e. $g^*(x) = 1 \Leftrightarrow x \in D_1^* \Leftrightarrow \eta(x) > 1/2$.

THEOREM

The Bayes decision minimizes $r(g, x)$ $\forall x$, and so optimal. The minimum is $r(g^*, x) = \min(\eta_0(x), \eta_1(x))$.

(optimal) global risk of g^*: *Bayes risk/Bayes error*

$$R^* \stackrel{\text{def}}{=} R(g^*) = \mathbb{E}\left[\min(\eta_0(X), \eta_1(X))\right] = \mathbb{E}\left[\min(\eta(X), 1 - \eta(X))\right].$$

OTHER FORMULAS FOR BAYES RISK

$$R^* = \inf_{g:\mathcal{X}\to\{0,1\}} \mathbb{P}\left(g(X) \neq Y\right) = \frac{1}{2} - \frac{1}{2}\mathbb{E}\left[|2\eta(X) - 1|\right].$$

If X has density f:

$$R^* = \int \min(\eta(x), 1-\eta(x)) f(x) dx = \int \min\left((1-p)f_0(x), pf_1(x)\right) dx,$$

where $p = \mathbb{P}\left(Y = 1\right)$, $1 - p$ are the *class probabilities*, f_i is the class-conditional density of X given $Y = i$. If f_0 and f_1 are nonoverlapping, i.e., $\int f_0 f_1 = 0 \Rightarrow R^* = 0$.
If $p = 1/2$

$$R^* = \frac{1}{2}\int \min\left(f_0(x), f_1(x)\right) dx = \frac{1}{2} - \frac{1}{4}\int |f_1(x) - f_0(x)| dx,$$

i.e. is related to the L_1 distance between f_0,f_1.

OUTLINE

APPROXIMATION OF BAYES DECISION

- η is typically unknown.
- Assume that η_i can be estimated by some $\tilde{\eta}_i : \mathcal{X} \to [0, 1]$.
- Bayes decision: $(\eta_0, \eta_1) \Rightarrow g^*$.
  Analogy: $(\tilde{\eta}_0, \tilde{\eta}_1) \Rightarrow \tilde{g}$ defines a *plug-in decision*:

  $$\tilde{g}(x) = j \Rightarrow \tilde{\eta}_j(x) = \max(\tilde{\eta}_0(x), \tilde{\eta}_1(x))$$

  (if $\tilde{\eta}_0(x) = \tilde{\eta}_1(x)$, choose arbitrarily, e.g., 0.)
- Expectation: $\tilde{\eta}_i$'s are good estimates $\Rightarrow$ $\tilde{g}$'s error $\approx$ g^*'s error (always $\geq$). Diff. of their risks $\leq$ estimation errors of $\tilde{\eta}_i$'s

APPROXIMATION OF BAYES DECISION

- η is typically unknown.
- Assume that η_i can be estimated by some $\tilde{\eta}_i : \mathcal{X} \to [0, 1]$.
- Bayes decision: $(\eta_0, \eta_1) \Rightarrow g^*$.
  Analogy: $(\tilde{\eta}_0, \tilde{\eta}_1) \Rightarrow \tilde{g}$ defines a *plug-in decision*:

  $$\tilde{g}(x) = j \Rightarrow \tilde{\eta}_j(x) = \max(\tilde{\eta}_0(x), \tilde{\eta}_1(x))$$

  (if $\tilde{\eta}_0(x) = \tilde{\eta}_1(x)$, choose arbitrarily, e.g., 0.)
- Expectation: $\tilde{\eta}_i$'s are good estimates $\Rightarrow$ $\tilde{g}$'s error $\approx$ g^*'s error (always $\geq$). Diff. of their risks $\leq$ estimation errors of $\tilde{\eta}_i$'s

APPROXIMATION OF BAYES DECISION

- η is typically unknown.
- Assume that η_i can be estimated by some $\tilde{\eta}_i : \mathcal{X} \to [0, 1]$.
- Bayes decision: $(\eta_0, \eta_1) \Rightarrow g^*$.
  Analogy: $(\tilde{\eta}_0, \tilde{\eta}_1) \Rightarrow \tilde{g}$ defines a *plug-in decision*:

$$\tilde{g}(x) = j \Rightarrow \tilde{\eta}_j(x) = \max(\tilde{\eta}_0(x), \tilde{\eta}_1(x))$$

  (if $\tilde{\eta}_0(x) = \tilde{\eta}_1(x)$, choose arbitrarily, e.g., 0.)
- Expectation: $\tilde{\eta}_i$'s are good estimates $\Rightarrow$ $\tilde{g}$'s error $\approx g^*$'s error (always $\geq$). Diff. of their risks $\leq$ estimation errors of $\tilde{\eta}_i$'s

APPROXIMATION OF BAYES DECISION

- η is typically unknown.
- Assume that η_i can be estimated by some $\tilde{\eta}_i : \mathcal{X} \to [0, 1]$.
- Bayes decision: $(\eta_0, \eta_1) \Rightarrow g^*$.
  Analogy: $(\tilde{\eta}_0, \tilde{\eta}_1) \Rightarrow \tilde{g}$ defines a *plug-in decision*:

  $$\tilde{g}(x) = j \Rightarrow \tilde{\eta}_j(x) = \max(\tilde{\eta}_0(x), \tilde{\eta}_1(x))$$

  (if $\tilde{\eta}_0(x) = \tilde{\eta}_1(x)$, choose arbitrarily, e.g., 0.)
- Expectation: $\tilde{\eta}_i$'s are good estimates $\Rightarrow$ $\tilde{g}$'s error $\approx$ g^*'s error (always $\geq$). Diff. of their risks $\leq$ estimation errors of $\tilde{\eta}_i$'s

APPROXIMATION OF BAYES DECISION 2

THEOREM

For $i = 0, 1$ let $\tilde{\eta}_i : \mathcal{X} \to [0, 1]$ be estimate of η_i and $\tilde{g}$ be the plug-in decision function defined by $(\tilde{\eta}_0, \tilde{\eta}_1)$. Then

$$r(\tilde{g}, x) - r(g^*, x) \leq \mathbb{I}_{\{\tilde{g}(x) \neq g^*(x)\}} \sum_{i \in \{0,1\}} |\tilde{\eta}_i(x) - \eta_i(x)|$$

and $$R(\tilde{g}) - R^* \leq$$

$$\mathbb{E}\left[\mathbb{I}_{\{\tilde{g}(X) \neq g^*(X)\}} \sum_{i \in \{0,1\}} |\tilde{\eta}_i(x) - \eta_i(x)|\right] \leq \mathbb{E}\left[\sum_{i \in \{0,1\}} |\tilde{\eta}_i(X) - \eta_i(X)|\right].$$

- If $1 - \tilde{\eta}_0 = \tilde{\eta}_1 \stackrel{\text{def}}{=} \tilde{\eta}$ then $r(\tilde{g}, x) - r(g^*, x) = \mathbb{I}_{\{\tilde{g}(x) \neq g^*(x)\}} |1 - 2\eta(x)| \leq 2\mathbb{I}_{\{\tilde{g}(x) \neq g^*(x)\}} |\tilde{\eta}(x) - \eta(x)|$ and $R(\tilde{g}) - R^* = \mathbb{E}\left[\mathbb{I}_{\{\tilde{g}(X) \neq g^*(X)\}} |1 - 2\eta(x)|\right] \leq 2\mathbb{E}\left[|\tilde{\eta}(X) - \eta(X)|\right]$.
- Good η estimate $\Rightarrow$ good decision function

APPROXIMATION OF BAYES DECISION 2

THEOREM

For $i = 0, 1$ let $\tilde{\eta}_i : \mathcal{X} \to [0, 1]$ be estimate of η_i and $\tilde{g}$ be the plug-in decision function defined by $(\tilde{\eta}_0, \tilde{\eta}_1)$. Then

$$r(\tilde{g}, x) - r(g^*, x) \leq \mathbb{I}_{\{\tilde{g}(x) \neq g^*(x)\}} \sum_{i \in \{0,1\}} |\tilde{\eta}_i(x) - \eta_i(x)|$$

and $\quad R(\tilde{g}) - R^* \leq$

$$\mathbb{E}\left[\mathbb{I}_{\{\tilde{g}(X) \neq g^*(X)\}} \sum_{i \in \{0,1\}} |\tilde{\eta}_i(x) - \eta_i(x)|\right] \leq \mathbb{E}\left[\sum_{i \in \{0,1\}} |\tilde{\eta}_i(X) - \eta_i(X)|\right].$$

- If $1 - \tilde{\eta}_0 = \tilde{\eta}_1 \stackrel{\text{def}}{=} \tilde{\eta}$ then $\quad r(\tilde{g}, x) - r(g^*, x) = \mathbb{I}_{\{\tilde{g}(x) \neq g^*(x)\}} |1 - 2\eta(x)| \leq 2\mathbb{I}_{\{\tilde{g}(x) \neq g^*(x)\}} |\tilde{\eta}(x) - \eta(x)|$ and $R(\tilde{g}) - R^* = \mathbb{E}\left[\mathbb{I}_{\{\tilde{g}(X) \neq g^*(X)\}} |1 - 2\eta(x)|\right] \leq 2\mathbb{E}\left[|\tilde{\eta}(X) - \eta(X)|\right].$
- Good η estimate $\Rightarrow$ good decision function

APPROXIMATION OF BAYES DECISION 2

THEOREM

For $i = 0,1$ let $\tilde{\eta}_i : \mathcal{X} \to [0,1]$ be estimate of η_i and $\tilde{g}$ be the plug-in decision function defined by $(\tilde{\eta}_0, \tilde{\eta}_1)$. Then

$$r(\tilde{g},x) - r(g^*,x) \leq \mathbb{I}_{\{\tilde{g}(x)\neq g^*(x)\}} \sum\nolimits_{i\in\{0,1\}} |\tilde{\eta}_i(x) - \eta_i(x)|$$

and $R(\tilde{g}) - R^* \leq$

$$\mathbb{E}\left[\mathbb{I}_{\{\tilde{g}(X)\neq g^*(X)\}} \sum_{i\in\{0,1\}} |\tilde{\eta}_i(x) - \eta_i(x)|\right] \leq \mathbb{E}\left[\sum_{i\in\{0,1\}} |\tilde{\eta}_i(X) - \eta_i(X)|\right].$$

- If $1 - \tilde{\eta}_0 = \tilde{\eta}_1 \stackrel{\text{def}}{=} \tilde{\eta}$ then $r(\tilde{g},x) - r(g^*,x) = \mathbb{I}_{\{\tilde{g}(x)\neq g^*(x)\}} |1 - 2\eta(x)| \leq 2\mathbb{I}_{\{\tilde{g}(x)\neq g^*(x)\}} |\tilde{\eta}(x) - \eta(x)|$ and $R(\tilde{g}) - R^* = \mathbb{E}\left[\mathbb{I}_{\{\tilde{g}(X)\neq g^*(X)\}} |1-2\eta(x)|\right] \leq 2\mathbb{E}\left[|\tilde{\eta}(X) - \eta(X)|\right].$
- Good η estimate $\Rightarrow$ good decision function

APPROXIMATION OF BAYES DECISION 3

If X has a density, f_0, f_1 are estimated by densities $\tilde{f}_0, \tilde{f}_1$, and p, $1-p$ are estimated by $\tilde{p}_1$, $\tilde{p}_0$, respectively, then for the plug-in decision function

$$g(x) = \begin{cases} 1 & \text{if } \tilde{p}_1 \tilde{f}_1(x) > \tilde{p}_0 \tilde{f}_0(x) \\ 0 & \text{otherwise,} \end{cases}$$

$$\begin{aligned} & R(g) - R^* \\ & \quad \leq \int_{\mathcal{X}} |(1-p) f_0(x) - \tilde{p}_0 \tilde{f}_0(x)| dx + \int_{\mathcal{X}} |p f_1(x) - \tilde{p}_1 \tilde{f}_1(x)| dx. \end{aligned}$$

OUTLINE

1. INTRO: DECISION, SUPERVISED (PASSIVE) LEARNING
   - Bayes decision
   - Approximation of Bayes decision
   - Sample based classification
   - No rate - Slow rate of convergence
   - Resticted class - Empirical risk minimization

SAMPLE BASED CLASSIFICATION

η is unknown. **Assumption**: we have i.i.d. *data* (*sample*, *observations*) $D_n = ((X_1, Y_1),\ldots,(X_n, Y_n)) \sim \nu$ from experiment or experts (strong, but can be extended for slightly dependent data).

An approximating classifier g_n is constructed based on D_n (Y is guessed by $g_n(X; D_n)$). So $g_n : \mathcal{X} \times \{\mathcal{X} \times \{0, 1\}\}^n \longrightarrow \{0, 1\}$.

$\Rightarrow$ *Classification*, *Pattern Recognition*, or *(Supervised) Learning (with a teacher)*

Performance of g_n is measured by conditional error prob. $R_n \stackrel{\text{def}}{=} R(g_n) = \mathbb{P}\left(g_n(X; D_n) \neq Y | D_n\right)$, it depends on the data $\Rightarrow$ random variable! But bounded: $R_n \in [0, 1]$

A sequence $\{g_n, n \geq 1\}$ is a *(discrimination) rule*.

CONSISTENT RULES

When is $\{g_n\}$ good?

DEFINITION

$\{g_n\}$ is *(weakly) consistent* if $R_n \to R^*$ in probability (equivalently, $\lim_{n\to\infty} \mathbb{E}[R_n] = R^*$), and *strongly consistent* if $R_n \to R^*$ a.s., i.e. $\mathbb{P}(R_n \to R^*) = 1$. If a rule is (weekly/strongly) consistent for all ν on $\mathcal{X} \times \{0,1\}$, then it is *universally (weekly/strongly) consistent*.

Consistency assures that taking more samples suffices to roughly reconstruct needed aspects of μ (actually, g^*).
1^{st} universal consistency proof: Stone'77, k-NN rule ($k(n) \to \infty$ and $k(n)/n \to 0$). k-NN: $g_n(x)$ takes majority vote over Y_i's in the subset of k pairs from D_n for which X_i is nearest to x. Since then many rules have been shown to be universally consistent.
For most well-behaved $\{g_n\}$ (e.g. k-NN), weak and strong consistency are equivalent $\Leftarrow$ concentration inequalities

HOEFFDING INEQUALITY

See lecture01_ucb.pdf Sec.4 p.15!

OUTLINE

NO RATE - SLOW RATE OF CONVERGENCE

How good can $\{g_n\}$ be? Convergence $\Leftrightarrow$ explicit inequality $R_n \geq R^*$. Desire: bounds on $\mathbb{E}[R_n] - R^*$ and $\mathbb{P}(R_n - R^* > \epsilon)$ *Rate of convergence* But! Such bound has to depend on ν. E.g:

THEOREM

$\forall \epsilon > 0$, *n, and* g_n, $\exists (X, Y) \sim \nu$ *with* $R^* = 0$ *s.t.* $\mathbb{E}[R_n] \geq 1/2 - \epsilon$.

THEOREM

Let $\{a_n\}$ *be a real sequence with* $a_n \to 0$, $1/16 \geq a_1 \geq a_2 \geq \ldots > 0$. $\forall \{g_n\}, \exists (X, Y) \sim \nu$ *with* $R^* = 0$, *s.t.* $\forall n \, \mathbb{E}[R_n] \geq a_n$.

THEOREM

$\forall \{g_n\}, \epsilon, \liminf_{n \to \infty} \sup_{\text{all } \nu \text{ with } R^* < 1/2 - \epsilon} \mathbb{P}(R_n - R^* > \epsilon) > 0$.

Universal convergence rate guarantees do not exist. They must involve certain subclasses of distributions of (X, Y).

NO RATE - SLOW RATE OF CONVERGENCE

How good can $\{g_n\}$ be? Convergence $\Leftrightarrow$ explicit inequality $R_n \geq R^*$. Desire: bounds on $\mathbb{E}[R_n] - R^*$ and $\mathbb{P}(R_n - R^* > \epsilon)$ *Rate of convergence* But! Such bound has to depend on ν. E.g:

THEOREM

$\forall \epsilon > 0$, n, and g_n, $\exists (X, Y) \sim \nu$ with $R^* = 0$ s.t. $\mathbb{E}[R_n] \geq 1/2 - \epsilon$.

THEOREM

Let $\{a_n\}$ be a real sequence with $a_n \to 0$, $1/16 \geq a_1 \geq a_2 \geq \ldots > 0$. $\forall \{g_n\}$, $\exists (X, Y) \sim \nu$ with $R^* = 0$, s.t. $\forall n\, \mathbb{E}[R_n] \geq a_n$.

THEOREM

$\forall \{g_n\}, \epsilon$, $\liminf_{n \to \infty} \sup_{\text{all } \nu \text{ with } R^* < 1/2 - \epsilon} \mathbb{P}(R_n - R^* > \epsilon) > 0$.

Universal convergence rate guarantees do not exist. They must involve certain subclasses of distributions of (X, Y).

NO RATE - SLOW RATE OF CONVERGENCE

How good can $\{g_n\}$ be? Convergence $\Leftrightarrow$ explicit inequality $R_n \geq R^*$. Desire: bounds on $\mathbb{E}[R_n] - R^*$ and $\mathbb{P}(R_n - R^* > \epsilon)$ *Rate of convergence* But! Such bound has to depend on ν. E.g:

THEOREM

$\forall \epsilon > 0$, *n, and* g_n, $\exists (X, Y) \sim \nu$ *with* $R^* = 0$ *s.t.* $\mathbb{E}[R_n] \geq 1/2 - \epsilon$.

THEOREM

Let $\{a_n\}$ *be a real sequence with* $a_n \to 0$, $1/16 \geq a_1 \geq a_2 \geq \ldots > 0$. $\forall \{g_n\}$, $\exists (X, Y) \sim \nu$ *with* $R^* = 0$, *s.t.* $\forall n\, \mathbb{E}[R_n] \geq a_n$.

THEOREM

$\forall \{g_n\}, \epsilon$, $\liminf_{n \to \infty} \sup_{\text{all } \nu \text{ with } R^* < 1/2 - \epsilon} \mathbb{P}(R_n - R^* > \epsilon) > 0$.

Universal convergence rate guarantees do not exist. They must involve certain subclasses of distributions of (X, Y).

OUTLINE

1. INTRO: DECISION, SUPERVISED (PASSIVE) LEARNING
   - Bayes decision
   - Approximation of Bayes decision
   - Sample based classification
   - No rate - Slow rate of convergence
   - Resticted class - Empirical risk minimization

RESTICTED CLASS - EMPIRICAL RISK MINIMIZATION

Change the setting: limit the classifiers to class $\mathcal{F}$ such as, e.g., neural networks with k node in 1 hidden layes. Then picking g_n from $\mathcal{F}$, $R_m \geq R_{\mathcal{F}} \stackrel{\text{def}}{=} \inf_{g \in \mathcal{F}} R(g)$. Typically, $R_{\mathcal{F}} > R^*$.
How to find a good $g_n \in \mathcal{F}$? Pick a g_n^* with minimal estimated error, e.g. *minimize empirical risk* over $\mathcal{F}$:

$$\widehat{R}_n(g) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} I_{\{g(X_i) \neq Y_i\}}.$$

(Algorithmic complexity?! - not here)
$R(g_n^*) - R_{\mathcal{F}} \geq 0$, but expected to become small. Can we give convergence rate on it for such classes? Yes! Distribution free bounds, 1$^{\text{st}}$ by Vapnik & Chervonenkis, 1971.
$R(g_n^*) - R^* = (R(g_n^*) - R_{\mathcal{F}}) + (R_{\mathcal{F}} - R^*)$ decomposition estimation error + approximation error $\Rightarrow$ trade-off!

FINITE CLASS

THEOREM

Let $|\mathcal{F}| < \infty$ *and* $R_{\mathcal{F}} = 0$*. Then* $\forall n, \epsilon > 0$,

$$\mathbb{P}\left(R(g_n^*) > \epsilon\right) \leq |\mathcal{F}| e^{-n\epsilon} \quad \text{and} \quad \mathbb{E}\left[R(g_n^*)\right] \leq \frac{\log(e|\mathcal{F}|)}{n}.$$

PROOF. $R_{\mathcal{F}} = 0 \Rightarrow \exists g \in \mathcal{F}$: $R(g) = 0 \Rightarrow \widehat{R}_n(g) = 0 \Rightarrow \widehat{R}_n(g_n^*) = 0$ a.s.

$$\mathbb{P}\left(R(g_n^*) > \epsilon\right) \leq \mathbb{P}\left(\max_{g \in \mathcal{F}: \widehat{R}_n(g)=0} R(g) > \epsilon\right)$$

$$= \mathbb{E}\left[\mathbb{I}_{\left\{\max_{g \in \mathcal{F}: \widehat{R}_n(g)=0} R(g) > \epsilon\right\}}\right] = \mathbb{E}\left[\max_{g \in \mathcal{F}} \mathbb{I}_{\left\{\widehat{R}_n(g)=0\right\}} \mathbb{I}_{\{R(g) > \epsilon\}}\right]$$

$$\leq \sum_{g \in \mathcal{F}: R(g) > \epsilon} \mathbb{P}\left(\widehat{R}_n(g) = 0\right) \leq |\mathcal{F}|(1-\epsilon)^n \leq |\mathcal{F}| e^{-n\epsilon}$$

$(\mathbb{P}\left(\nexists (X_i, Y_i) \in \{(x, y) : g(x) \neq y\}\right) < (1-\epsilon)^n$ if $\mathbb{P}\left(g(X) \neq Y\right) > \epsilon)$

FINITE CLASS PROOF - CONT.

$\forall u > 0$,

$$\begin{aligned}
\mathbb{E}\left[R(g_n^*)\right] &= \int_0^\infty \mathbb{P}\left(R(g_n^*) > \epsilon\right) d\epsilon \leq u + \int_u^\infty \mathbb{P}\left(R(g_n^*) > \epsilon\right) d\epsilon \\
&\leq u + |\mathcal{F}| \int_u^\infty e^{-n\epsilon} d\epsilon = u + \frac{|\mathcal{F}|}{n} e^{-nu}.
\end{aligned}$$

Set $u = \log|\mathcal{F}|/n \Rightarrow$ bound $\log(e|\mathcal{F}|)/n$. $\square$

REFERENCES I

Devroye, L., Györfi, L., and Lugosi, G. (1996).
A Probabilistic Theory of Pattern Recognition.
Applications of Mathematics: Stochastic Modelling and Applied Probability. Springer-Verlag New York.