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
1. **Intro: Decision, Supervised (Passive) Learning**

- Bayes decision
- Approximation of Bayes decision
- Sample based classification
- No rate - Slow rate of convergence
- Restricted 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
   - Restricted 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. $\nu$
- $g: \mathcal{X} \rightarrow \mathcal{Y}$ 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): $R(g) \overset{\text{def}}{=} \mathbb{P}(Y \neq g(X))$
- $g$’s minimizing $R(g)$: optimal
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. $\nu$

- $g : \mathcal{X} \rightarrow \mathcal{Y}$ 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

- Performance of $g$ is measured by its error probability (global risk): $R(g) \overset{\text{def}}{=} \mathbb{P}(Y \neq g(X))$

- $g$'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. $\nu$
- $g : \mathcal{X} \rightarrow \mathcal{Y}$ 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
  - Performance of $g$ is measured by its error probability (global risk): $R(g) \overset{\text{def}}{=} \mathbb{P}(Y \neq g(X))$
  - $g$’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. $\nu$
- $g : \mathcal{X} \rightarrow \mathcal{Y}$ 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): $R(g) \overset{\text{def}}{=} \mathbb{P}(Y \neq g(X))$
  - $g$'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. $\nu$
- $g : \mathcal{X} \to \mathcal{Y}$ 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 ⇒
- Performance of $g$ is measured by its error probability (global risk): $R(g) \overset{\text{def}}{=} \mathbb{P}(Y \neq g(X))$
- $g$’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. $\nu$
- $g : \mathcal{X} \rightarrow \mathcal{Y}$ 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): $R(g) \overset{\text{def}}{=} \mathbb{P}(Y \neq g(X))$
- $g$’s minimizing $R(g)$: optimal
For $i = 0, 1$: $\{Y = i\}$: $i^{th}$ hypothesis. $Y$ a posteriori distribution is given by $\eta_i(x) \overset{\text{def}}{=} \mathbb{P}(Y = i|X = x)$ 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: $1 - \eta_0(x) = \eta_1(x) = \mathbb{E}[Y|X = x] \overset{\text{def}}{=} \eta(x)$ (regression or a posteriori probability function).

If $X \sim \mu$, $\nu$ may be given, e.g., by $(\mu, \eta)$. $\forall C_0, C_1 \subseteq \mathcal{X}$

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

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

$r(g, x) \overset{\text{def}}{=} \mathbb{P}(Y \neq g(X) | X = x)$: local risk function

$\mathbb{E}[r(g, X)] = R(g)$ and

\[
    r(g, x) = \mathbb{I}_{\{g(x) = 1\}} \eta_0(x) + \mathbb{I}_{\{g(x) = 0\}} \eta_1(x) \\
    = 1 - \mathbb{I}_{\{x \in D_0\}} \eta_0(x) - \mathbb{I}_{\{x \in D_1\}} \eta_1(x).
\]

Minimized by $g$ which puts $\forall x$ into $D_j$ with the greater $\eta_j(x)$. 
1. **Intro: Decision, Supervised (Passive) Learning**
   - Bayes decision
   - Approximation of Bayes decision
   - Sample based classification
   - No rate - Slow rate of convergence
   - Restricted class - Empirical risk minimization
Bayes decision

Let \( \{D_j^*\} \) be s.t. \( \forall x \)

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

\( g^* \) 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. \( g^*(x) = 1 \iff x \in D_1^* \iff \eta(x) > 1/2. \)

**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

\[
R^* \overset{\text{def}}{=} R(g^*) = \mathbb{E} [\min(\eta_0(X), \eta_1(X))] = \mathbb{E} [\min(\eta(X), 1 - \eta(X))].
\]
Let \( \{D_j^*\} \) be s.t. \( \forall x \)
\[
x \in D_j^* \iff (\eta_j(x) > \eta_{1-j}(x) \text{ v. } j = 0, \eta_0(x) = \eta_1(x))
\]
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. \( g^*(x) = 1 \iff x \in D_1^* \iff \eta(x) > 1/2. \)

**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
\[
R^* \overset{\text{def}}{=} R(g^*) = \mathbb{E} \left[ \min(\eta_0(X), \eta_1(X)) \right] = \mathbb{E} \left[ \min(\eta(X), 1 - \eta(X)) \right].
\]
Let \( \{D_j^*\} \) be s.t. \( \forall x \)
\[
x \in D_j^* \iff (\eta_j(x) > \eta_{1-j}(x)) \text{ v. } j = 0, \eta_0(x) = \eta_1(x)
\]
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. \( g^*(x) = 1 \iff x \in D_1^* \iff \eta(x) > 1/2. \)

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

Let \( \{D_j^*\} \) be s.t. \( \forall x \)
\[
x \in D_j^* \iff (\eta_j(x) > \eta_{1-j}(x) \text{ v. } j = 0, \eta_0(x) = \eta_1(x))
\]
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. \( g^*(x) = 1 \iff x \in D_1^* \iff \eta(x) > 1/2 \).

**Theorem**

The Bayes decision minimizes \( r(g, x) \forall x, \) and so optimal. The minimum is \( r(g^*, x) = \min(\eta_0(x), \eta_1(x)) \).
INTRO: DECISION, SUPERVISED (PASSIVE) LEARNING

Bayes decision

- Let \( \{D_j^*\} \) be s.t. \( \forall x \)
  \[
  x \in D_j^* \iff (\eta_j(x) > \eta_{1-j}(x)) \text{ v. } j = 0, \eta_0(x) = \eta_1(x)
  \]
  \( g^* \) 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. \( g^*(x) = 1 \iff x \in D_1^* \iff \eta(x) > 1/2. \)

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

\[
R^* \overset{\text{def}}{=} R(g^*) = \mathbb{E} \left[ \min(\eta_0(X), \eta_1(X)) \right] = \mathbb{E} \left[ \min(\eta(X), 1 - \eta(X)) \right].
\]
Other formulas for Bayes risk

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

If \( X \) has density \( f \):

\[ R^* = \int \min(\eta(x), 1-\eta(x)) f(x) dx = \int \min((1-p)f_0(x), pf_1(x)) dx, \]

where \( p = \mathbb{P}(Y = 1), 1 - p \) 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 \( p = 1/2 \)

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

i.e. is related to the \( L_1 \) distance between \( f_0, f_1 \).
**Intro: Decision, Supervised (Passive) Learning**

**Outline**

1. **Intro: Decision, Supervised (Passive) Learning**
   - Bayes decision
   - Approximation of Bayes decision
   - Sample based classification
   - No rate - Slow rate of convergence
   - Restricted class - Empirical risk minimization
Approximation of Bayes Decision

- $\eta$ is typically unknown.
- Assume that $\eta_i$ can be estimated by some $\tilde{\eta}_i : \mathcal{X} \rightarrow [0, 1]$.
- 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
\textbf{Approximation of Bayes decision}

- $\eta$ is typically unknown.
- Assume that $\eta_i$ 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

- $\eta$ is typically unknown.
- Assume that $\eta_i$ can be estimated by some $\tilde{\eta}_i : \mathcal{X} \rightarrow [0, 1]$.
- 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

- $\eta$ is typically unknown.
- Assume that $\eta_i$ can be estimated by some $\tilde{\eta}_i : \mathcal{X} \rightarrow [0, 1]$.
- 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
For $i = 0, 1$ let $\tilde{\eta}_i : \mathcal{X} \to [0, 1]$ be estimate of $\eta_i$ 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 $1 - \tilde{\eta}_0 = \tilde{\eta}_1 \overset{\text{def}}{=} \tilde{\eta}$ 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 $\eta$ estimate $\implies$ good decision function.
**THEOREM**

For \( i = 0, 1 \) let \( \tilde{\eta}_i : \mathcal{X} \to [0, 1] \) be estimate of \( \eta_i \) 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 \( 1 - \tilde{\eta}_0 = \tilde{\eta}_1 \overset{\text{def}}{=} \tilde{\eta} \) 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 \( \eta \) 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 $\eta_i$ 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 $1 - \tilde{\eta}_0 = \tilde{\eta}_1 \overset{\text{def}}{=} \tilde{\eta}$ 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 $\eta$ estimate $\Rightarrow$ good decision function
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

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

$$R(g) - R^* \leq \int_{X} |(1 - p)f_0(x) - \tilde{p}_0 \tilde{f}_0(x)| \, dx + \int_{X} |pf_1(x) - \tilde{p}_1 \tilde{f}_1(x)| \, dx.$$
INTRO: DECISION, SUPERVISED (PASSIVE) LEARNING

OUTLINE

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

\( \eta \) 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 \( g_n : \mathcal{X} \times \{\mathcal{X} \times \{0, 1\}\}^n \rightarrow \{0, 1\} \).

⇒ **Classification, Pattern Recognition, or (Supervised) Learning (with a teacher)**

Performance of \( g_n \) is measured by conditional error prob.

\[ R_n \stackrel{\text{def}}{=} R(g_n) = \mathbb{P}(g_n(X; D_n) \neq Y|D_n), \]

it depends on the data ⇒ random variable! But bounded: \( R_n \in [0, 1] \)

A sequence \( \{g_n, n \geq 1\} \) is a (discrimination) rule.
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 \( \nu \) on \( X \times \{0, 1\} \), then it is universally (weekly/strongly) consistent.

*Consistency* assures that taking more samples suffices to roughly reconstruct needed aspects of \( \mu \) (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 \( \iff \) concentration inequalities.
Hoeffding inequality

See lecture01_ucb.pdf Sec.4 p.15!
**OUTLINE**

1. Intro: Decision, Supervised (Passive) Learning
   - Bayes decision
   - Approximation of Bayes decision
   - Sample based classification
   - No rate - Slow rate of convergence
   - Restricted class - Empirical risk minimization
**No rate - Slow rate of convergence**

How good can \( \{g_n\} \) be? Convergence ⇔ explicit inequality \( R_n \geq R^* \). Desire: bounds on \( \mathbb{E} [R_n] - R^* \) and \( \mathbb{P} (R_n - R^* > \varepsilon) \)

**Rate of convergence** But! Such bound has to depend on \( \nu \). E.g:

**Theorem**

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

**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 \text{ with } R^* = 0, \text{ s.t. } \forall n \mathbb{E} [R_n] \geq a_n.

**Theorem**

\[
\forall \{g_n\}, \varepsilon, \lim \inf_{n \to \infty} \sup_{\text{all } \nu \text{ with } R^* < 1/2 - \varepsilon} \mathbb{P} (R_n - R^* > \varepsilon) > 0.
\]

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 \( E[R_n] - R^* \) and \( P(R_n - R^* > \epsilon) \)

Rate of convergence But! Such bound has to depend on \( \nu \). E.g:

**Theorem**

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

**Theorem**

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

**Theorem**

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

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 ⇔ 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 \( \nu \). E.g:

**Theorem**

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

**Theorem**

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

**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 \((X, Y)\).

1 I N T R O : D E C I S I O N , S U P E R V I S E D ( P A S S I V E ) L E A R N I N G
- Bayes decision
- Approximation of Bayes decision
- Sample based classification
- No rate - Slow rate of convergence
- Restricted 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 layer. Then picking $g_n$ from $\mathcal{F}$, $R_m \geq R_{\mathcal{F}} \overset{\text{def}}{=} \inf_{g \in \mathcal{F}} R(g)$. 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 $\mathcal{F}$:

$$\hat{R}_n(g) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} I\{g(X_i) \neq Y_i\}.$$  

(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, 1st 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!
Let $|\mathcal{F}| < \infty$ and $R_\mathcal{F} = 0$. Then $\forall n, \epsilon > 0$,

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

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

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

$$
= \mathbb{E}\left[\mathbb{1}\{\max_{g \in \mathcal{F}: \hat{R}_n(g) = 0} R(g) > \epsilon\}\right] = \mathbb{E}\left[\max_{g \in \mathcal{F}} \mathbb{1}\{\hat{R}_n(g) = 0\}\mathbb{1}\{R(g) > \epsilon\}\right]
$$

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

$(\mathbb{P}(\exists (X_i, Y_i) \in \{(x, y): g(x) \neq y\}) < (1 - \epsilon)^n \text{ if } \mathbb{P}(g(X) \neq Y) > \epsilon)$
\[ \forall u > 0, \]
\[ \mathbb{E}[R(g_n^*)] = \int_0^\infty \mathbb{P}(R(g_n^*) > \epsilon) \, d\epsilon \leq u + \int_u^\infty \mathbb{P}(R(g_n^*) > \epsilon) \, d\epsilon \]
\[ \leq u + |\mathcal{F}| \int_u^\infty e^{-n\epsilon} \, d\epsilon = u + \frac{|\mathcal{F}|}{n} e^{-nu}. \]

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

*A Probabilistic Theory of Pattern Recognition.*