Aula 30 - Teoria do Aprendizado (Parte II)

João B. Florindo

Instituto de Matemática, Estatística e Computação Científica
Universidade Estadual de Campinas - Brasil
florindo@unicamp.br

Outline

- Seja a classe finita $\mathcal{H} = \{h_1, \cdots, h_k\}$.
- Este é um conjunto de k funções que mapeiam de $\mathcal{X}$ para $\{0, 1\}$.
- ERM seleciona entre essas k funções qual tem menor erro de treino.
- Queremos dar garantias sobre o erro de generalização de h. Para isto:
  1. Mostramos que $\hat{\varepsilon}(h)$ é uma estimativa confiável de $\varepsilon(h)$ para todo h.
  2. Mostramos que isso implica em um limitante superior para $\varepsilon(h)$.

- Seja a classe finita $\mathcal{H} = \{h_1, \cdots, h_k\}$.
- Este é um conjunto de k funções que mapeiam de $\mathcal{X}$ para $\{0, 1\}$.
- ERM seleciona entre essas k funções qual tem menor erro de treino.
- Queremos dar garantias sobre o erro de generalização de h. Para isto:
  1. Mostramos que $\hat{\varepsilon}(h)$ é uma estimativa confiável de $\varepsilon(h)$ para todo h.
  2. Mostramos que isso implica em um limitante superior para $\varepsilon(h)$.

- Seja a classe finita $\mathcal{H} = \{h_1, \cdots, h_k\}$.
- Este é um conjunto de k funções que mapeiam de $\mathcal{X}$ para $\{0, 1\}$.
- ERM seleciona entre essas k funções qual tem menor erro de treino.
- Queremos dar garantias sobre o erro de generalização de h. Para isto:
  1. Mostramos que $\hat{\varepsilon}(h)$ é uma estimativa confiável de $\varepsilon(h)$ para todo h.
  2. Mostramos que isso implica em um limitante superior para $\varepsilon(h)$.

- Seja a classe finita $\mathcal{H} = \{h_1, \cdots, h_k\}$.
- Este é um conjunto de k funções que mapeiam de $\mathcal{X}$ para $\{0, 1\}$.
- ERM seleciona entre essas k funções qual tem menor erro de treino.
- Queremos dar garantias sobre o erro de generalização de h. Para isto:
  1. Mostramos que $\hat{\varepsilon}(h)$ é uma estimativa confiável de $\varepsilon(h)$ para todo h.
  2. Mostramos que isso implica em um limitante superior para $\varepsilon(h)$.

- Seja a classe finita $\mathcal{H} = \{h_1, \cdots, h_k\}$.
- Este é um conjunto de k funções que mapeiam de $\mathcal{X}$ para $\{0, 1\}$.
- ERM seleciona entre essas k funções qual tem menor erro de treino.
- Queremos dar garantias sobre o erro de generalização de h. Para isto:
  1. Mostramos que $\hat{\varepsilon}(h)$ é uma estimativa confiável de $\varepsilon(h)$ para todo h.
  2. Mostramos que isso implica em um limitante superior para $\varepsilon(h)$.

- Seja a classe finita $\mathcal{H} = \{h_1, \cdots, h_k\}$.
- Este é um conjunto de k funções que mapeiam de $\mathcal{X}$ para $\{0, 1\}$.
- ERM seleciona entre essas k funções qual tem menor erro de treino.
- Queremos dar garantias sobre o erro de generalização de h. Para isto:
  1. Mostramos que $\hat{\varepsilon}(h)$ é uma estimativa confiável de $\varepsilon(h)$ para todo h.
  2. Mostramos que isso implica em um limitante superior para $\varepsilon(h)$.

- Vamos partir de uma hipótese $h_i \in \mathcal{H}$ fixa.
- Definimos uma variável aleatória de Bernoulli Z.
- Amostramos $(x, y) \sim \mathcal{D}$ e fazemos

$$Z = \mathbb{1}\{h_i(x) \neq y\}.$$

- Similarmente, definimos

$$Z_j = \mathbb{1}\{h_i(x^{(j)}) \neq y^{(j)}\}.$$

- Como o conjunto de treino é amostrado iid de $\mathcal{D}$, Z e Z_j têm a mesma distribuição.

- Vamos partir de uma hipótese $h_i \in \mathcal{H}$ fixa.
- Definimos uma variável aleatória de Bernoulli Z.
- Amostramos $(x, y) \sim \mathcal{D}$ e fazemos

$$Z = \mathbb{1}\{h_i(x) \neq y\}.$$

- Similarmente, definimos

$$Z_j = \mathbb{1}\{h_i(x^{(j)}) \neq y^{(j)}\}.$$

- Como o conjunto de treino é amostrado iid de $\mathcal{D}$, Z e Z_j têm a mesma distribuição.

- Vamos partir de uma hipótese $h_i \in \mathcal{H}$ fixa.
- Definimos uma variável aleatória de Bernoulli Z.
- Amostramos $(x, y) \sim \mathcal{D}$ e fazemos

$$Z = \mathbb{1}\{h_i(x) \neq y\}.$$

- Similarmente, definimos

$$Z_j = \mathbb{1}\{h_i(x^{(j)}) \neq y^{(j)}\}.$$

- Como o conjunto de treino é amostrado iid de $\mathcal{D}$, Z e Z_j têm a mesma distribuição.

- Vamos partir de uma hipótese $h_i \in \mathcal{H}$ fixa.
- Definimos uma variável aleatória de Bernoulli Z.
- Amostramos $(x, y) \sim \mathcal{D}$ e fazemos

$$Z = \mathbb{1}\{h_i(x) \neq y\}.$$

- Similarmente, definimos

$$Z_j = \mathbb{1}\{h_i(x^{(j)}) \neq y^{(j)}\}.$$

- Como o conjunto de treino é amostrado iid de $\mathcal{D}$, Z e Z_j têm a mesma distribuição.

- Vamos partir de uma hipótese $h_i \in \mathcal{H}$ fixa.
- Definimos uma variável aleatória de Bernoulli Z.
- Amostramos $(x, y) \sim \mathcal{D}$ e fazemos

$$Z = \mathbb{1}\{h_i(x) \neq y\}.$$

- Similarmente, definimos

$$Z_j = \mathbb{1}\{h_i(x^{(j)}) \neq y^{(j)}\}.$$

- Como o conjunto de treino é amostrado iid de $\mathcal{D}$, Z e Z_j têm a mesma distribuição.

- A probabilidade de classificação incorreta para um exemplo aleatório, i.e., $\varepsilon(h)$ é o valor esperado de Z (e Z_j).

- Já para o erro de treinamento temos

$$\hat{\varepsilon}(h_i) = \frac{1}{n} \sum_{j=1}^{n} Z_j.$$

- Portanto, $\hat{\varepsilon}(h_i)$ é a média de n variáveis aleatórias amostradas iid de uma distribuição de Bernoulli com média $\varepsilon(h_i)$.

- Podemos aplicar então a desigualdade de Hoeffding:

$$P(|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) \leq 2 \exp(-2\gamma^2 n).$$

- A probabilidade de classificação incorreta para um exemplo aleatório, i.e., $\varepsilon(h)$ é o valor esperado de Z (e Z_j).

- Já para o erro de treinamento temos

$$\hat{\varepsilon}(h_i) = \frac{1}{n} \sum_{j=1}^{n} Z_j.$$

- Portanto, $\hat{\varepsilon}(h_i)$ é a média de n variáveis aleatórias amostradas iid de uma distribuição de Bernoulli com média $\varepsilon(h_i)$.

- Podemos aplicar então a desigualdade de Hoeffding:

$$P(|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) \leq 2 \exp(-2\gamma^2 n).$$

- A probabilidade de classificação incorreta para um exemplo aleatório, i.e., $\varepsilon(h)$ é o valor esperado de Z (e Z_j).

- Já para o erro de treinamento temos

$$\hat{\varepsilon}(h_i) = \frac{1}{n} \sum_{j=1}^{n} Z_j.$$

- Portanto, $\hat{\varepsilon}(h_i)$ é a média de n variáveis aleatórias amostradas iid de uma distribuição de Bernoulli com média $\varepsilon(h_i)$.

- Podemos aplicar então a desigualdade de Hoeffding:

$$P(|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) \leq 2 \exp(-2\gamma^2 n).$$

- A probabilidade de classificação incorreta para um exemplo aleatório, i.e., $\varepsilon(h)$ é o valor esperado de Z (e Z_j).

- Já para o erro de treinamento temos

$$\hat{\varepsilon}(h_i) = \frac{1}{n} \sum_{j=1}^{n} Z_j.$$

- Portanto, $\hat{\varepsilon}(h_i)$ é a média de n variáveis aleatórias amostradas iid de uma distribuição de Bernoulli com média $\varepsilon(h_i)$.

- Podemos aplicar então a desigualdade de Hoeffding:

$$P(|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) \leq 2 \exp(-2\gamma^2 n).$$

- Então, para nossa h_i específica, o erro de generalização está próximo do de treino com alta probabilidade se n é grande.
- Mas queremos mostrar isso para *toda* $h \in \mathcal{H}$.
- Seja A_i o evento $|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma$.
- Já vimos que para um A_i específico:

$$P(A_i) \leq 2\exp(-2\gamma^2 n).$$

- Então, para nossa h_i específica, o erro de generalização está próximo do de treino com alta probabilidade se n é grande.
- Mas queremos mostrar isso para *toda* $h \in \mathcal{H}$.
- Seja A_i o evento $|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma$.
- Já vimos que para um A_i específico:

$$P(A_i) \leq 2\exp(-2\gamma^2 n).$$

- Então, para nossa h_i específica, o erro de generalização está próximo do de treino com alta probabilidade se n é grande.
- Mas queremos mostrar isso para *toda* $h \in \mathcal{H}$.
- Seja A_i o evento $|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma$.
- Já vimos que para um A_i específico:

$$P(A_i) \leq 2\exp(-2\gamma^2 n).$$

- Então, para nossa h_i específica, o erro de generalização está próximo do de treino com alta probabilidade se n é grande.
- Mas queremos mostrar isso para *toda* $h \in \mathcal{H}$.
- Seja A_i o evento $|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma$.
- Já vimos que para um A_i específico:

$$P(A_i) \leq 2\exp(-2\gamma^2 n).$$

- Pelo limitante da união:

$$\begin{aligned} P(\exists h \in \mathcal{H} : |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) &= P(A_1 \cup \cdots \cup A_k) \\ &\leq \sum_{i=1}^{k} P(A_i) \\ &\leq \sum_{i=1}^{k} 2\exp(-2\gamma^2 n) \\ &= 2k\exp(-2\gamma^2 n). \end{aligned}$$

- Subtraindo 1 de ambos os lados:

$$\begin{aligned} P(\nexists h \in \mathcal{H} : |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) &= P(\forall h \in \mathcal{H} : |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| \leq \gamma) \\ &\geq 1 - 2k\exp(-2\gamma^2 n). \end{aligned}$$

- Pelo limitante da união:

$$\begin{aligned} P(\exists h \in \mathcal{H} : |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) &= P(A_1 \cup \cdots \cup A_k) \\ &\leq \sum_{i=1}^{k} P(A_i) \\ &\leq \sum_{i=1}^{k} 2\exp(-2\gamma^2 n) \\ &= 2k\exp(-2\gamma^2 n). \end{aligned}$$

- Subtraindo 1 de ambos os lados:

$$\begin{aligned} P(\nexists h \in \mathcal{H} : |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) &= P(\forall h \in \mathcal{H} : |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| \leq \gamma) \\ &\geq 1 - 2k\exp(-2\gamma^2 n). \end{aligned}$$

- Então, com probabilidade no mínimo $1 - 2k\exp(-2\gamma^2 n)$, temos que $\varepsilon(h)$ está a uma distância no máximo γ de $\hat{\varepsilon}(h)$, para todo $h \in \mathcal{H}$.

- Este é um resultado de *convergência uniforme* porque é um limitante que vale simultaneamente para todo $h \in \mathcal{H}$.

- Então, com probabilidade no mínimo $1 - 2k\exp(-2\gamma^2 n)$, temos que $\varepsilon(h)$ está a uma distância no máximo γ de $\hat{\varepsilon}(h)$, para todo $h \in \mathcal{H}$.

- Este é um resultado de *convergência uniforme* porque é um limitante que vale simultaneamente para todo $h \in \mathcal{H}$.

- Até agora limitamos a probabilidade de erro, dados n e γ.

- Mas e se nos forem dados γ e algum $\delta > 0$, quão grande deve ser n para que o erro de treino fique a uma distância no máximo γ do erro de generalização com probabilidade no mínimo $1 - \delta$?

- Para responder, fazemos $\delta = 2k \exp(-2\gamma^2 n)$ e resolvemos para n, tal que

$$n \geq \frac{1}{2\gamma^2} \log \frac{2k}{\delta}$$

e então, com probabilidade ao menos $1 - \delta$, temos $|\varepsilon(h) - \hat{\varepsilon}(h)| \leq \gamma$ para todo $h \in \mathcal{H}$.

- Ou ainda: a probabilidade de $|\varepsilon(h) - \hat{\varepsilon}(h)| > \gamma$ para algum $h \in \mathcal{H}$ é no máximo δ.

- Até agora limitamos a probabilidade de erro, dados n e γ.

- Mas e se nos forem dados γ e algum $\delta > 0$, quão grande deve ser n para que o erro de treino fique a uma distância no máximo γ do erro de generalização com probabilidade no mínimo $1 - \delta$?

- Para responder, fazemos $\delta = 2k \exp(-2\gamma^2 n)$ e resolvemos para n, tal que

$$n \geq \frac{1}{2\gamma^2} \log \frac{2k}{\delta}$$

e então, com probabilidade ao menos $1 - \delta$, temos $|\varepsilon(h) - \hat{\varepsilon}(h)| \leq \gamma$ para todo $h \in \mathcal{H}$.

- Ou ainda: a probabilidade de $|\varepsilon(h) - \hat{\varepsilon}(h)| > \gamma$ para algum $h \in \mathcal{H}$ é no máximo δ.

- Até agora limitamos a probabilidade de erro, dados n e γ.

- Mas e se nos forem dados γ e algum $\delta > 0$, quão grande deve ser n para que o erro de treino fique a uma distância no máximo γ do erro de generalização com probabilidade no mínimo $1 - \delta$?

- Para responder, fazemos $\delta = 2k \exp(-2\gamma^2 n)$ e resolvemos para n, tal que

$$n \geq \frac{1}{2\gamma^2} \log \frac{2k}{\delta}$$

e então, com probabilidade ao menos $1 - \delta$, temos $|\varepsilon(h) - \hat{\varepsilon}(h)| \leq \gamma$ para todo $h \in \mathcal{H}$.

- Ou ainda: a probabilidade de $|\varepsilon(h) - \hat{\varepsilon}(h)| > \gamma$ para algum $h \in \mathcal{H}$ é no máximo δ.

- Até agora limitamos a probabilidade de erro, dados n e γ.

- Mas e se nos forem dados γ e algum $\delta > 0$, quão grande deve ser n para que o erro de treino fique a uma distância no máximo γ do erro de generalização com probabilidade no mínimo $1 - \delta$?

- Para responder, fazemos $\delta = 2k \exp(-2\gamma^2 n)$ e resolvemos para n, tal que

$$n \geq \frac{1}{2\gamma^2} \log \frac{2k}{\delta}$$

  e então, com probabilidade ao menos $1 - \delta$, temos $|\varepsilon(h) - \hat{\varepsilon}(h)| \leq \gamma$ para todo $h \in \mathcal{H}$.

- Ou ainda: a probabilidade de $|\varepsilon(h) - \hat{\varepsilon}(h)| > \gamma$ para algum $h \in \mathcal{H}$ é no máximo δ.

- Este tamanho de treino n necessário para certo nível de performance é chamado de **complexidade amostral**.
- Note que aqui, o n necessário é apenas *logarítmico* em k (número de hipóteses em $\mathcal{H}$).
- Usaremos isso mais tarde.

- Este tamanho de treino n necessário para certo nível de performance é chamado de **complexidade amostral**.
- Note que aqui, o n necessário é apenas *logarítmico* em k (número de hipóteses em $\mathcal{H}$).
- Usaremos isso mais tarde.

- Este tamanho de treino n necessário para certo nível de performance é chamado de **complexidade amostral**.
- Note que aqui, o n necessário é apenas *logarítmico* em k (número de hipóteses em $\mathcal{H}$).
- Usaremos isso mais tarde.

- Podemos também fixar n e δ e resolver para γ.
- Então, com probabilidade $1 - \delta$ temos para todo $h \in \mathcal{H}$:

$$|\varepsilon(h) - \hat{\varepsilon}(h)| \leq \sqrt{\frac{1}{2n} \log \frac{2k}{\delta}}.$$

- Podemos também fixar n e δ e resolver para γ.
- Então, com probabilidade $1 - \delta$ temos para todo $h \in \mathcal{H}$:

$$|\varepsilon(h) - \hat{\varepsilon}(h)| \leq \sqrt{\frac{1}{2n} \log \frac{2k}{\delta}}.$$

- Vamos focar agora na generalização da melhor hipótese $\hat{h} = \text{argmin}_{h \in \mathcal{H}} \hat{\varepsilon}(H)$, que nosso algoritmo de aprendizado encontra.
- Seja $h^* = \text{argmin}_{h \in \mathcal{H}} \varepsilon(H)$ a melhor hipótese possível em $\mathcal{H}$.
- Assumimos convergência uniforme, i.e., $|\varepsilon(h) - \hat{\varepsilon}(h)| \leq \gamma$ para todo $h \in \mathcal{H}$.
- Então teremos:

$$\begin{aligned} \varepsilon(\hat{h}) &\leq \hat{\varepsilon}(\hat{h}) + \gamma \\ &\leq \hat{\varepsilon}(h^*) + \gamma \\ &\leq \varepsilon(h^*) + 2\gamma. \end{aligned}$$

- Vamos focar agora na generalização da melhor hipótese $\hat{h} = \text{argmin}_{h \in \mathcal{H}} \hat{\varepsilon}(H)$, que nosso algoritmo de aprendizado encontra.
- Seja $h^* = \text{argmin}_{h \in \mathcal{H}} \varepsilon(H)$ a melhor hipótese possível em $\mathcal{H}$.
- Assumimos convergência uniforme, i.e., $|\varepsilon(h) - \hat{\varepsilon}(h)| \leq \gamma$ para todo $h \in \mathcal{H}$.
- Então teremos:

$$\begin{aligned} \varepsilon(\hat{h}) &\leq \hat{\varepsilon}(\hat{h}) + \gamma \\ &\leq \hat{\varepsilon}(h^*) + \gamma \\ &\leq \varepsilon(h^*) + 2\gamma. \end{aligned}$$

- Vamos focar agora na generalização da melhor hipótese $\hat{h} = \text{argmin}_{h \in \mathcal{H}} \hat{\varepsilon}(H)$, que nosso algoritmo de aprendizado encontra.
- Seja $h^* = \text{argmin}_{h \in \mathcal{H}} \varepsilon(H)$ a melhor hipótese possível em $\mathcal{H}$.
- Assumimos convergência uniforme, i.e., $|\varepsilon(h) - \hat{\varepsilon}(h)| \leq \gamma$ para todo $h \in \mathcal{H}$.
- Então teremos:

$$\begin{aligned} \varepsilon(\hat{h}) &\leq \hat{\varepsilon}(\hat{h}) + \gamma \\ &\leq \hat{\varepsilon}(h^*) + \gamma \\ &\leq \varepsilon(h^*) + 2\gamma. \end{aligned}$$

- Vamos focar agora na generalização da melhor hipótese $\hat{h} = \text{argmin}_{h \in \mathcal{H}} \hat{\varepsilon}(H)$, que nosso algoritmo de aprendizado encontra.
- Seja $h^* = \text{argmin}_{h \in \mathcal{H}} \varepsilon(H)$ a melhor hipótese possível em $\mathcal{H}$.
- Assumimos convergência uniforme, i.e., $|\varepsilon(h) - \hat{\varepsilon}(h)| \leq \gamma$ para todo $h \in \mathcal{H}$.
- Então teremos:

$$\begin{aligned} \varepsilon(\hat{h}) &\leq \hat{\varepsilon}(\hat{h}) + \gamma \\ &\leq \hat{\varepsilon}(h^*) + \gamma \\ &\leq \varepsilon(h^*) + 2\gamma. \end{aligned}$$

- A 1ª linha vem de $|\varepsilon(\hat{h}) - \hat{\varepsilon}(\hat{h})| \leq \gamma$, devido à convergência uniforme.
- A 2ª vem de que $\hat{h}$ minimiza $\hat{\varepsilon}(h)$ e portanto $\hat{\varepsilon}(\hat{h}) \leq \hat{\varepsilon}(h)$ para todo h e, em particular, $\hat{\varepsilon}(\hat{h}) \leq \hat{\varepsilon}(h^*)$.
- A 3ª vem de aplicarmos a convergência uniforme novamente, de modo que $\hat{\varepsilon}(h^*) \leq \varepsilon(h^*) + \gamma$.
- Então, em condições de convergência uniforme, temos que o erro de generalização de $\hat{h}$ é no máximo 2γ pior do que a melhor hipótese possível em $\mathcal{H}$.

- A 1ª linha vem de $|\varepsilon(\hat{h}) - \hat{\varepsilon}(\hat{h})| \leq \gamma$, devido à convergência uniforme.
- A 2ª vem de que $\hat{h}$ minimiza $\hat{\varepsilon}(h)$ e portanto $\hat{\varepsilon}(\hat{h}) \leq \hat{\varepsilon}(h)$ para todo h e, em particular, $\hat{\varepsilon}(\hat{h}) \leq \hat{\varepsilon}(h^*)$.
- A 3ª vem de aplicarmos a convergência uniforme novamente, de modo que $\hat{\varepsilon}(h^*) \leq \varepsilon(h^*) + \gamma$.
- Então, em condições de convergência uniforme, temos que o erro de generalização de $\hat{h}$ é no máximo 2γ pior do que a melhor hipótese possível em $\mathcal{H}$.

- A 1ª linha vem de $|\varepsilon(\hat{h}) - \hat{\varepsilon}(\hat{h})| \leq \gamma$, devido à convergência uniforme.
- A 2ª vem de que $\hat{h}$ minimiza $\hat{\varepsilon}(h)$ e portanto $\hat{\varepsilon}(\hat{h}) \leq \hat{\varepsilon}(h)$ para todo h e, em particular, $\hat{\varepsilon}(\hat{h}) \leq \hat{\varepsilon}(h^*)$.
- A 3ª vem de aplicarmos a convergência uniforme novamente, de modo que $\hat{\varepsilon}(h^*) \leq \varepsilon(h^*) + \gamma$.
- Então, em condições de convergência uniforme, temos que o erro de generalização de $\hat{h}$ é no máximo 2γ pior do que a melhor hipótese possível em $\mathcal{H}$.

- A 1ª linha vem de $|\varepsilon(\hat{h}) - \hat{\varepsilon}(\hat{h})| \leq \gamma$, devido à convergência uniforme.
- A 2ª vem de que $\hat{h}$ minimiza $\hat{\varepsilon}(h)$ e portanto $\hat{\varepsilon}(\hat{h}) \leq \hat{\varepsilon}(h)$ para todo h e, em particular, $\hat{\varepsilon}(\hat{h}) \leq \hat{\varepsilon}(h^*)$.
- A 3ª vem de aplicarmos a convergência uniforme novamente, de modo que $\hat{\varepsilon}(h^*) \leq \varepsilon(h^*) + \gamma$.
- Então, em condições de convergência uniforme, temos que o erro de generalização de $\hat{h}$ é no máximo 2γ pior do que a melhor hipótese possível em $\mathcal{H}$.

- Podemos juntar tudo em um teorema.

Teorema

Seja $|\mathcal{H}| = k$ e n, δ fixados. Então, com probabilidade ao menos $1 - \delta$, temos

$$\varepsilon(\hat{h}) \leq \left(\min_{h \in \mathcal{H}} \varepsilon(h) \right) + 2\sqrt{\frac{1}{2n} \log \frac{2k}{\delta}}.$$

- Este teorema dá a fundamentação teórica do dilema viés/variância.
- Se mudamos da classe de hipóteses $\mathcal{H}$ para uma classe maior $\mathcal{H}' \supseteq \mathcal{H}$, então o 1° termo $\min_{h \in \mathcal{H}}$ só pode decair pois buscamos o mínimo em um conjunto maior.
- Isso corresponde ao nosso "viés" decrescendo.
- Ao mesmo tempo, como k aumenta, o segundo termo $2\sqrt{\cdot}$, que corresponderia à "variância", aumenta.

- Podemos juntar tudo em um teorema.

Teorema

Seja $|\mathcal{H}| = k$ e n, δ fixados. Então, com probabilidade ao menos $1 - \delta$, temos

$$\varepsilon(\hat{h}) \leq \left(\min_{h \in \mathcal{H}} \varepsilon(h) \right) + 2\sqrt{\frac{1}{2n} \log \frac{2k}{\delta}}.$$

- Este teorema dá a fundamentação teórica do dilema viés/variância.
- Se mudamos da classe de hipóteses $\mathcal{H}$ para uma classe maior $\mathcal{H}' \supseteq \mathcal{H}$, então o 1° termo $\min_{h \in \mathcal{H}}$ só pode decair pois buscamos o mínimo em um conjunto maior.
- Isso corresponde ao nosso "viés" decrescendo.
- Ao mesmo tempo, como k aumenta, o segundo termo $2\sqrt{\cdot}$, que corresponderia à "variância", aumenta.

- Podemos juntar tudo em um teorema.

Teorema

Seja $|\mathcal{H}| = k$ e n, δ fixados. Então, com probabilidade ao menos $1 - \delta$, temos

$$\varepsilon(\hat{h}) \leq \left(\min_{h \in \mathcal{H}} \varepsilon(h) \right) + 2\sqrt{\frac{1}{2n} \log \frac{2k}{\delta}}.$$

- Este teorema dá a fundamentação teórica do dilema viés/variância.
- Se mudamos da classe de hipóteses $\mathcal{H}$ para uma classe maior $\mathcal{H}' \supseteq \mathcal{H}$, então o 1° termo $\min_{h \in \mathcal{H}}$ só pode decair pois buscamos o mínimo em um conjunto maior.
- Isso corresponde ao nosso "viés" decrescendo.
- Ao mesmo tempo, como k aumenta, o segundo termo $2\sqrt{\cdot}$, que corresponderia à "variância", aumenta.

- Podemos juntar tudo em um teorema.

Teorema

Seja $|\mathcal{H}| = k$ e n, δ fixados. Então, com probabilidade ao menos $1 - \delta$, temos

$$\varepsilon(\hat{h}) \leq \left(\min_{h \in \mathcal{H}} \varepsilon(h) \right) + 2\sqrt{\frac{1}{2n} \log \frac{2k}{\delta}}.$$

- Este teorema dá a fundamentação teórica do dilema viés/variância.
- Se mudamos da classe de hipóteses $\mathcal{H}$ para uma classe maior $\mathcal{H}' \supseteq \mathcal{H}$, então o 1º termo $\min_{h \in \mathcal{H}}$ só pode decair pois buscamos o mínimo em um conjunto maior.
- Isso corresponde ao nosso "viés" decrescendo.
- Ao mesmo tempo, como k aumenta, o segundo termo $2\sqrt{\cdot}$, que corresponderia à "variância", aumenta.

- Podemos juntar tudo em um teorema.

Teorema

Seja $|\mathcal{H}| = k$ e n, δ fixados. Então, com probabilidade ao menos $1 - \delta$, temos

$$\varepsilon(\hat{h}) \leq \left(\min_{h \in \mathcal{H}} \varepsilon(h) \right) + 2\sqrt{\frac{1}{2n} \log \frac{2k}{\delta}}.$$

- Este teorema dá a fundamentação teórica do dilema viés/variância.
- Se mudamos da classe de hipóteses $\mathcal{H}$ para uma classe maior $\mathcal{H}' \supseteq \mathcal{H}$, então o 1º termo $\min_{h \in \mathcal{H}}$ só pode decair pois buscamos o mínimo em um conjunto maior.
- Isso corresponde ao nosso "viés" decrescendo.
- Ao mesmo tempo, como k aumenta, o segundo termo $2\sqrt{\cdot}$, que corresponderia à "variância", aumenta.

- Podemos juntar tudo em um teorema.

Teorema

Seja $|\mathcal{H}| = k$ e n, δ fixados. Então, com probabilidade ao menos $1 - \delta$, temos

$$\varepsilon(\hat{h}) \leq \left(\min_{h \in \mathcal{H}} \varepsilon(h) \right) + 2\sqrt{\frac{1}{2n} \log \frac{2k}{\delta}}.$$

- Este teorema dá a fundamentação teórica do dilema viés/variância.
- Se mudamos da classe de hipóteses $\mathcal{H}$ para uma classe maior $\mathcal{H}' \supseteq \mathcal{H}$, então o 1º termo $\min_{h \in \mathcal{H}}$ só pode decair pois buscamos o mínimo em um conjunto maior.
- Isso corresponde ao nosso "viés" decrescendo.
- Ao mesmo tempo, como k aumenta, o segundo termo $2\sqrt{\cdot}$, que corresponderia à "variância", aumenta.

- Se fixarmos γ e δ e resolvermos para n, temos o seguinte limitante para a complexidade amostral:

Corolário

Seja $|\mathcal{H}| = k$ e δ, γ fixados. Então, para que tenhamos $\varepsilon(\hat{h}) \leq \min_{h \in \mathcal{H}} \varepsilon(h) + 2\gamma$ com probabilidade ao menos $1 - \delta$, basta que

$$n \geq \frac{1}{2\gamma^2} \log \frac{2k}{\delta}$$
$$= \mathcal{O}\left(\frac{1}{\gamma^2} \log \frac{k}{\delta}\right).$$

- Se fixarmos γ e δ e resolvermos para n, temos o seguinte limitante para a complexidade amostral:

Corolário

Seja $|\mathcal{H}| = k$ e δ, γ fixados. Então, para que tenhamos $\varepsilon(\hat{h}) \leq \min_{h \in \mathcal{H}} \varepsilon(h) + 2\gamma$ com probabilidade ao menos $1 - \delta$, basta que

$$n \geq \frac{1}{2\gamma^2} \log \frac{2k}{\delta}$$
$$= \mathcal{O}\left(\frac{1}{\gamma^2} \log \frac{k}{\delta}\right).$$

Outline

- Mostramos teoremas para o caso finito.
- Mas normalmente as classes de hipóteses são infinitas (ex. algoritmos parametrizados por números reais, como os classificadores lineares).
- Usaremos um argumento que não é o mais "correto", mas é intuitivo.
- Seja $\mathcal{H}$ parametrizado por d números reais.
- No computador, usamos ponto flutuante de precisão dupla (*double* em C).
- Nosso algoritmo de aprendizado é então parametrizado por $64d$ bits e temos $k = 2^{64d}$ hipóteses diferentes.

- Mostramos teoremas para o caso finito.
- Mas normalmente as classes de hipóteses são infinitas (ex. algoritmos parametrizados por números reais, como os classificadores lineares).
- Usaremos um argumento que não é o mais "correto", mas é intuitivo.
- Seja $\mathcal{H}$ parametrizado por d números reais.
- No computador, usamos ponto flutuante de precisão dupla (*double* em C).
- Nosso algoritmo de aprendizado é então parametrizado por $64d$ bits e temos $k = 2^{64d}$ hipóteses diferentes.

- Mostramos teoremas para o caso finito.
- Mas normalmente as classes de hipóteses são infinitas (ex. algoritmos parametrizados por números reais, como os classificadores lineares).
- Usaremos um argumento que não é o mais "correto", mas é intuitivo.
- Seja $\mathcal{H}$ parametrizado por d números reais.
- No computador, usamos ponto flutuante de precisão dupla (*double* em C).
- Nosso algoritmo de aprendizado é então parametrizado por $64d$ bits e temos $k = 2^{64d}$ hipóteses diferentes.

- Mostramos teoremas para o caso finito.
- Mas normalmente as classes de hipóteses são infinitas (ex. algoritmos parametrizados por números reais, como os classificadores lineares).
- Usaremos um argumento que não é o mais "correto", mas é intuitivo.
- Seja $\mathcal{H}$ parametrizado por d números reais.
- No computador, usamos ponto flutuante de precisão dupla (*double* em C).
- Nosso algoritmo de aprendizado é então parametrizado por $64d$ bits e temos $k = 2^{64d}$ hipóteses diferentes.

- Mostramos teoremas para o caso finito.
- Mas normalmente as classes de hipóteses são infinitas (ex. algoritmos parametrizados por números reais, como os classificadores lineares).
- Usaremos um argumento que não é o mais "correto", mas é intuitivo.
- Seja $\mathcal{H}$ parametrizado por d números reais.
- No computador, usamos ponto flutuante de precisão dupla (*double* em C).
- Nosso algoritmo de aprendizado é então parametrizado por $64d$ bits e temos $k = 2^{64d}$ hipóteses diferentes.

- Mostramos teoremas para o caso finito.
- Mas normalmente as classes de hipóteses são infinitas (ex. algoritmos parametrizados por números reais, como os classificadores lineares).
- Usaremos um argumento que não é o mais "correto", mas é intuitivo.
- Seja $\mathcal{H}$ parametrizado por d números reais.
- No computador, usamos ponto flutuante de precisão dupla (*double* em C).
- Nosso algoritmo de aprendizado é então parametrizado por $64d$ bits e temos $k = 2^{64d}$ hipóteses diferentes.

- Do corolário, temos que, para garantir $\varepsilon(\hat{h}) \leq \varepsilon(h^*) + 2\gamma$ com probabilidade no mínimo $1 - \delta$, basta que

$$\begin{aligned} n &\geq \mathcal{O}\left(\frac{1}{\gamma^2} \log \frac{2^{64d}}{\delta}\right) \\ &= \mathcal{O}\left(\frac{d}{\gamma^2} \log \frac{1}{\delta}\right) \\ &= \mathcal{O}_{\gamma,\delta}(d). \end{aligned}$$

- Ou seja, o número de exemplos de treino necessários é no máximo *linear* nos parâmetros do modelo.

- Embora o argumento não seja o mais preciso, esta conclusão é válida em geral.

- Do corolário, temos que, para garantir $\varepsilon(\hat{h}) \leq \varepsilon(h^*) + 2\gamma$ com probabilidade no mínimo $1 - \delta$, basta que

$$\begin{aligned} n &\geq \mathcal{O}\left(\frac{1}{\gamma^2} \log \frac{2^{64d}}{\delta}\right) \\ &= \mathcal{O}\left(\frac{d}{\gamma^2} \log \frac{1}{\delta}\right) \\ &= \mathcal{O}_{\gamma,\delta}(d). \end{aligned}$$

- Ou seja, o número de exemplos de treino necessários é no máximo *linear* nos parâmetros do modelo.

- Embora o argumento não seja o mais preciso, esta conclusão é válida em geral.

- Do corolário, temos que, para garantir $\varepsilon(\hat{h}) \leq \varepsilon(h^*) + 2\gamma$ com probabilidade no mínimo $1 - \delta$, basta que

$$\begin{aligned} n &\geq \mathcal{O}\left(\frac{1}{\gamma^2} \log \frac{2^{64d}}{\delta}\right) \\ &= \mathcal{O}\left(\frac{d}{\gamma^2} \log \frac{1}{\delta}\right) \\ &= \mathcal{O}_{\gamma,\delta}(d). \end{aligned}$$

- Ou seja, o número de exemplos de treino necessários é no máximo *linear* nos parâmetros do modelo.

- Embora o argumento não seja o mais preciso, esta conclusão é válida em geral.

- Note também que este resultado assume um algoritmo que usa minimização do risco empírico.

- Obter boas garantias teóricas para algoritmos não-ERM ainda é uma área ativa de pesquisa.

- Outro problema em nosso argumento é a parametrização de $\mathcal{H}$.

- Veja que a classe dos classificadores lineares em d dimensões pode ser escrita como $h_\theta(x) = \mathbb{1}\{\theta_0 + \theta_1 x_1 + \cdots + \theta_d x_d \geq 0\}$, com $d+1$ parâmetros $\theta_0, \cdots, \theta_d$.

- Mas também poderia ser escrita como $h_{u,v}(x) = \mathbb{1}\{(u_0^2 - v_0^2) + (u_1^2 - v_1^2)x_1 + \cdots + (u_d^2 - v_d^2)x_d \geq 0\}$, com $2(d+1)$ parâmetros u_i, v_i.

- Note também que este resultado assume um algoritmo que usa minimização do risco empírico.

- Obter boas garantias teóricas para algoritmos não-ERM ainda é uma área ativa de pesquisa.

- Outro problema em nosso argumento é a parametrização de $\mathcal{H}$.

- Veja que a classe dos classificadores lineares em d dimensões pode ser escrita como $h_\theta(x) = \mathbb{1}\{\theta_0 + \theta_1 x_1 + \cdots + \theta_d x_d \geq 0\}$, com $d+1$ parâmetros $\theta_0, \cdots, \theta_d$.

- Mas também poderia ser escrita como $h_{u,v}(x) = \mathbb{1}\{(u_0^2 - v_0^2) + (u_1^2 - v_1^2)x_1 + \cdots + (u_d^2 - v_d^2)x_d \geq 0\}$, com $2(d+1)$ parâmetros u_i, v_i.

- Note também que este resultado assume um algoritmo que usa minimização do risco empírico.
- Obter boas garantias teóricas para algoritmos não-ERM ainda é uma área ativa de pesquisa.
- Outro problema em nosso argumento é a parametrização de $\mathcal{H}$.
- Veja que a classe dos classificadores lineares em d dimensões pode ser escrita como $h_\theta(x) = \mathbb{1}\{\theta_0 + \theta_1 x_1 + \cdots + \theta_d x_d \geq 0\}$, com $d+1$ parâmetros $\theta_0, \cdots, \theta_d$.
- Mas também poderia ser escrita como $h_{u,v}(x) = \mathbb{1}\{(u_0^2 - v_0^2) + (u_1^2 - v_1^2)x_1 + \cdots + (u_d^2 - v_d^2)x_d \geq 0\}$, com $2(d+1)$ parâmetros u_i, v_i.

- Note também que este resultado assume um algoritmo que usa minimização do risco empírico.
- Obter boas garantias teóricas para algoritmos não-ERM ainda é uma área ativa de pesquisa.
- Outro problema em nosso argumento é a parametrização de $\mathcal{H}$.
- Veja que a classe dos classificadores lineares em d dimensões pode ser escrita como $h_\theta(x) = \mathbb{1}\{\theta_0 + \theta_1 x_1 + \cdots + \theta_d x_d \geq 0\}$, com $d+1$ parâmetros $\theta_0, \cdots, \theta_d$.
- Mas também poderia ser escrita como $h_{u,v}(x) = \mathbb{1}\{(u_0^2 - v_0^2) + (u_1^2 - v_1^2)x_1 + \cdots + (u_d^2 - v_d^2)x_d \geq 0\}$, com $2(d+1)$ parâmetros u_i, v_i.

- Note também que este resultado assume um algoritmo que usa minimização do risco empírico.
- Obter boas garantias teóricas para algoritmos não-ERM ainda é uma área ativa de pesquisa.
- Outro problema em nosso argumento é a parametrização de $\mathcal{H}$.
- Veja que a classe dos classificadores lineares em d dimensões pode ser escrita como $h_\theta(x) = \mathbb{1}\{\theta_0 + \theta_1 x_1 + \cdots + \theta_d x_d \geq 0\}$, com $d+1$ parâmetros $\theta_0, \cdots, \theta_d$.
- Mas também poderia ser escrita como $h_{u,v}(x) = \mathbb{1}\{(u_0^2 - v_0^2) + (u_1^2 - v_1^2)x_1 + \cdots + (u_d^2 - v_d^2)x_d \geq 0\}$, com $2(d+1)$ parâmetros u_i, v_i.

- Vamos derivar um argumento melhor definindo alguns conceitos novos.

- Seja $S = \{x^{(i)}, \cdots, x^{(\mathbf{D})}\}$ um conjunto de pontos $x^{(i)} \in \mathcal{X}$.

- Diz-se que $\mathcal{H}$ **shatters** ("destrói") S se $\mathcal{H}$ pode realizar qualquer rotulagem em S.

- I.e., para qualquer conjunto de rótulos $\{y^{(1)}, \cdots, y^{(\mathbf{D})}\}$, existe algum $h \in \mathcal{H}$ tal que $h(x^{(i)}) = y^{(i)}$ para todo $i = 1, \cdots, \mathbf{D}$.

- Vamos derivar um argumento melhor definindo alguns conceitos novos.
- Seja $S = \{x^{(i)}, \cdots, x^{(\mathbf{D})}\}$ um conjunto de pontos $x^{(i)} \in \mathcal{X}$.
- Diz-se que $\mathcal{H}$ **shatters** ("destrói") S se $\mathcal{H}$ pode realizar qualquer rotulagem em S.
- I.e., para qualquer conjunto de rótulos $\{y^{(1)}, \cdots, y^{(\mathbf{D})}\}$, existe algum $h \in \mathcal{H}$ tal que $h(x^{(i)}) = y^{(i)}$ para todo $i = 1, \cdots, \mathbf{D}$.

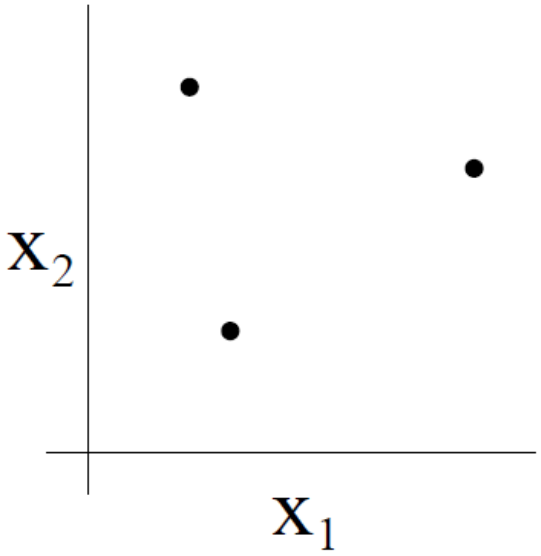
- Vamos derivar um argumento melhor definindo alguns conceitos novos.
- Seja $S = \{x^{(i)}, \cdots, x^{(\mathbf{D})}\}$ um conjunto de pontos $x^{(i)} \in \mathcal{X}$.
- Diz-se que $\mathcal{H}$ **shatters** ("destrói") S se $\mathcal{H}$ pode realizar qualquer rotulagem em S.
- I.e., para qualquer conjunto de rótulos $\{y^{(1)}, \cdots, y^{(\mathbf{D})}\}$, existe algum $h \in \mathcal{H}$ tal que $h(x^{(i)}) = y^{(i)}$ para todo $i = 1, \cdots, \mathbf{D}$.

- Vamos derivar um argumento melhor definindo alguns conceitos novos.
- Seja $S = \{x^{(i)}, \cdots, x^{(\mathbf{D})}\}$ um conjunto de pontos $x^{(i)} \in \mathcal{X}$.
- Diz-se que $\mathcal{H}$ **shatters** ("destrói") S se $\mathcal{H}$ pode realizar qualquer rotulagem em S.
- I.e., para qualquer conjunto de rótulos $\{y^{(1)}, \cdots, y^{(\mathbf{D})}\}$, existe algum $h \in \mathcal{H}$ tal que $h(x^{(i)}) = y^{(i)}$ para todo $i = 1, \cdots, \mathbf{D}$.

Dimensão VC

Dada uma classe de hipóteses $\mathcal{H}$, define-se a **dimensão de Vapnik-Chervonenkis** ($\mathrm{VC}(\mathcal{H})$) como o tamanho do maior conjunto que é *shattered* por $\mathcal{H}$.
Se $\mathcal{H}$ *shatters* conjuntos arbitrariamente grandes, $\mathrm{VC}(\mathcal{H}) = \infty$.

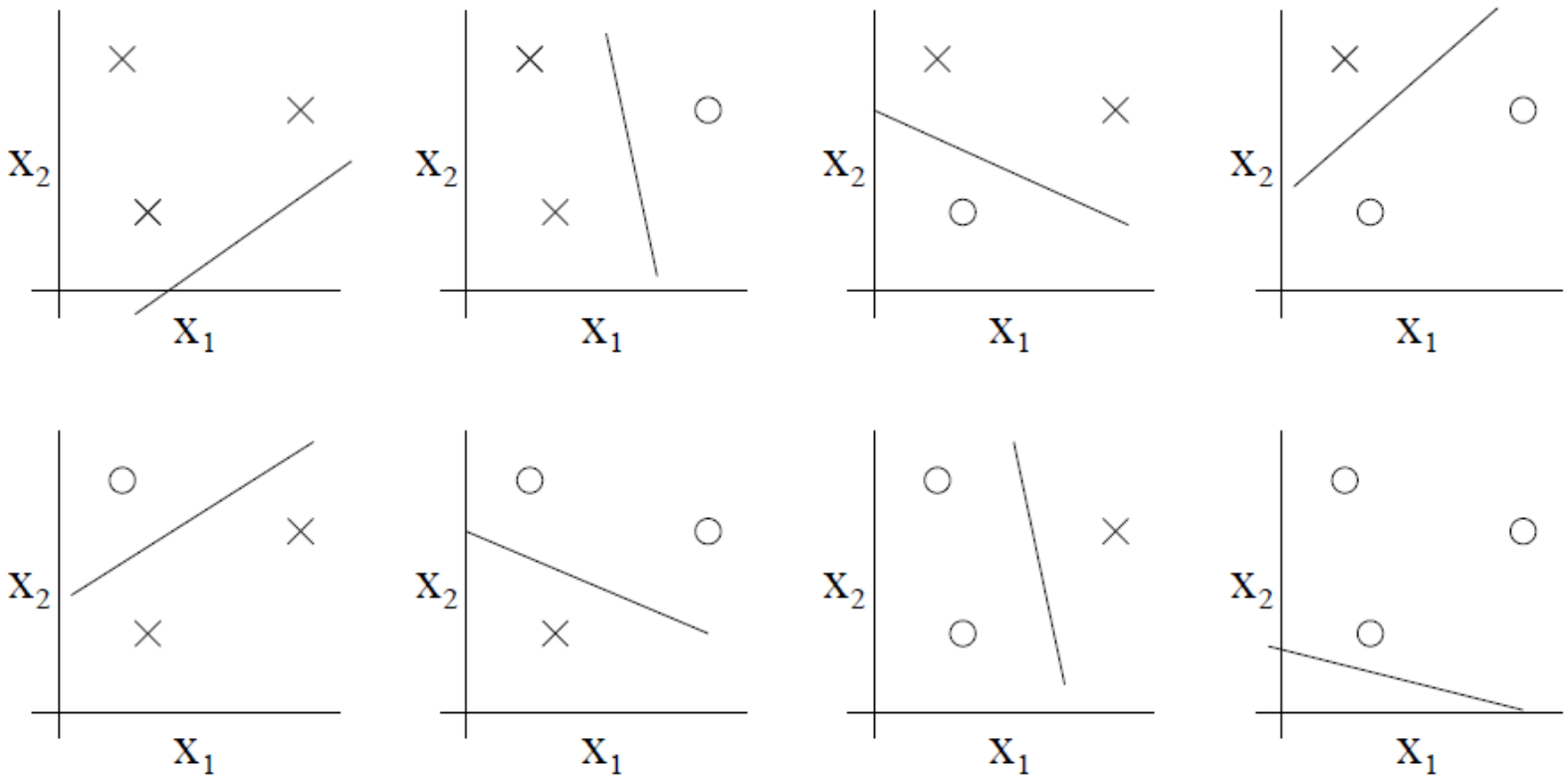
- Considere o conjunto de três pontos abaixo:



- Poderia a classe $\mathcal{H}$ dos classificadores lineares em duas dimensões ($h_\theta(x) = \mathbb{1}\{\theta_0 + \theta_1 x_1 + \theta_2 x_2 \geq 0\}$) *shatter* este conjunto?

- A resposta é SIM!

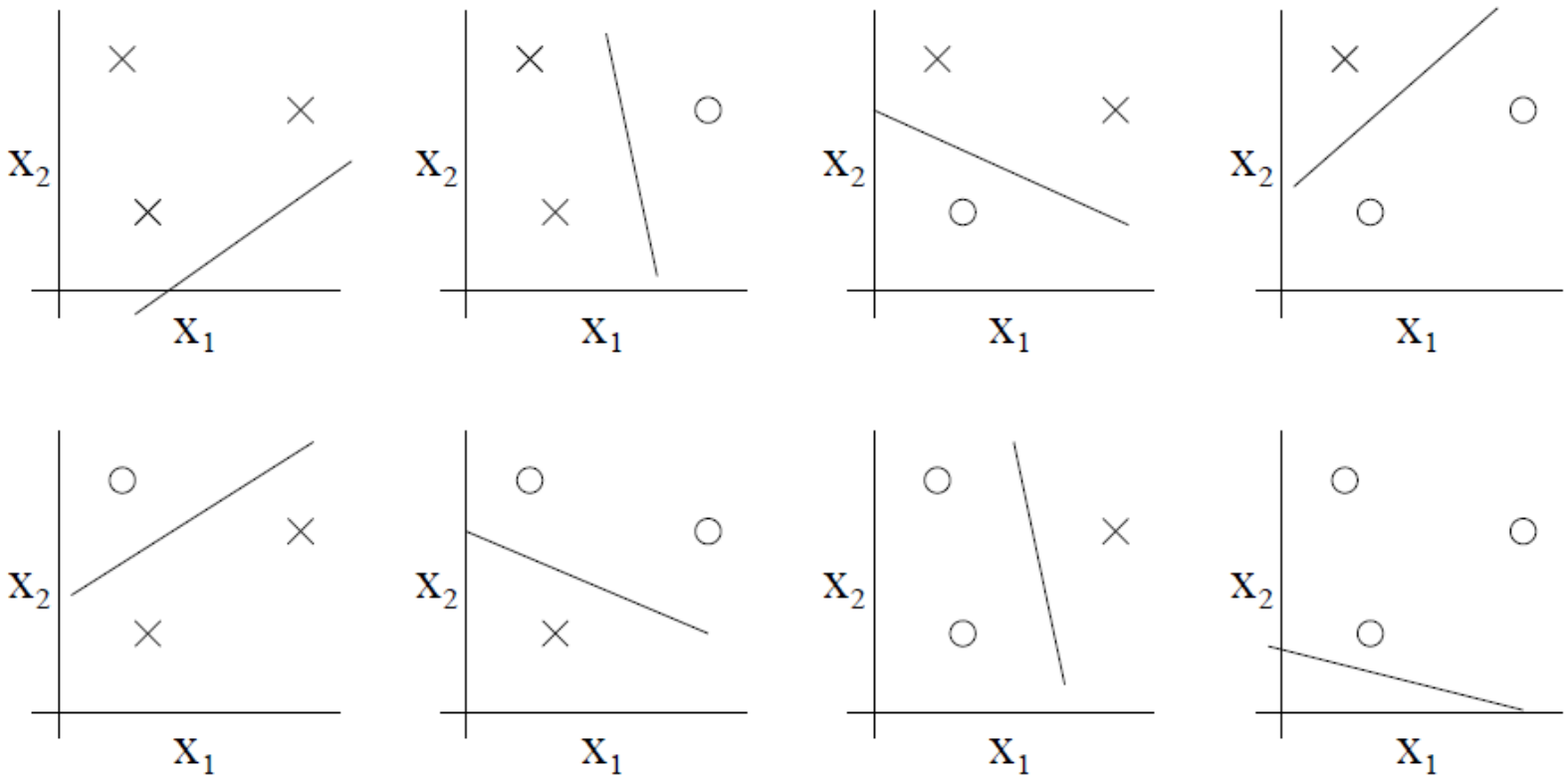
- Note que, para qualquer das 8 rotulagens possíveis para estes pontos, pode-se encontrar um classificador linear com "erro de treino zero":



- Poderia a classe $\mathcal{H}$ dos classificadores lineares em duas dimensões ($h_\theta(x) = \mathbb{1}\{\theta_0 + \theta_1 x_1 + \theta_2 x_2 \geq 0\}$) *shatter* este conjunto?

- A resposta é SIM!

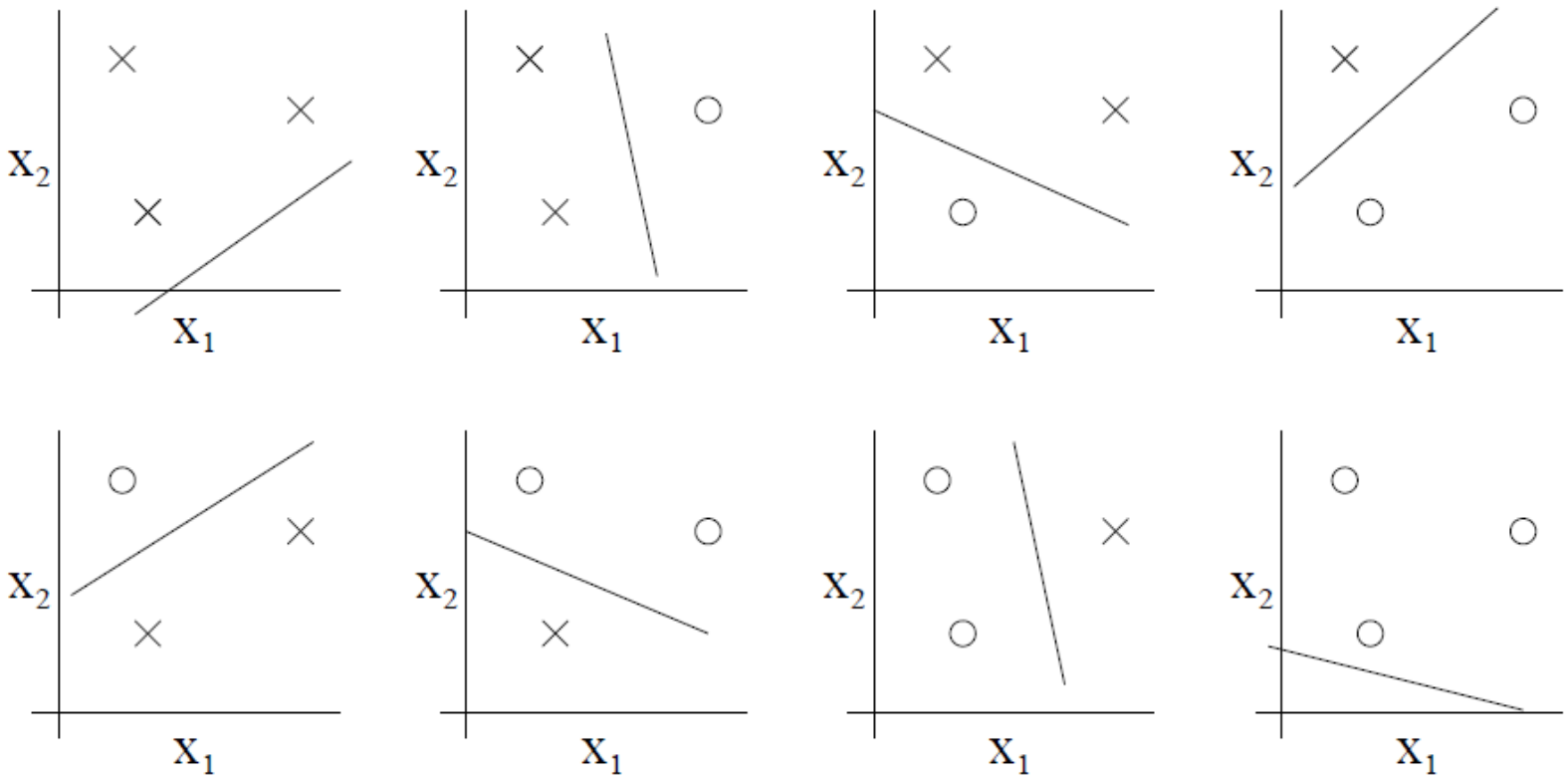
- Note que, para qualquer das 8 rotulagens possíveis para estes pontos, pode-se encontrar um classificador linear com "erro de treino zero":



- Poderia a classe $\mathcal{H}$ dos classificadores lineares em duas dimensões ($h_\theta(x) = \mathbb{1}\{\theta_0 + \theta_1 x_1 + \theta_2 x_2 \geq 0\}$) *shatter* este conjunto?

- A resposta é SIM!

- Note que, para qualquer das 8 rotulagens possíveis para estes pontos, pode-se encontrar um classificador linear com "erro de treino zero":

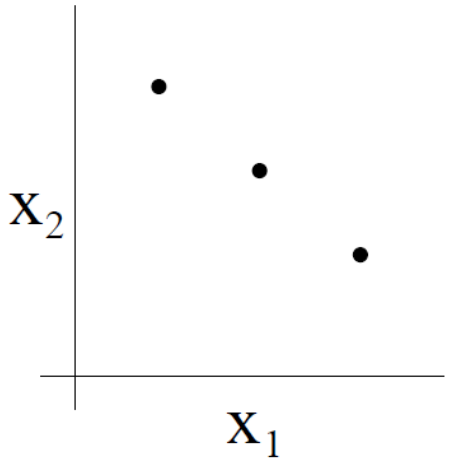


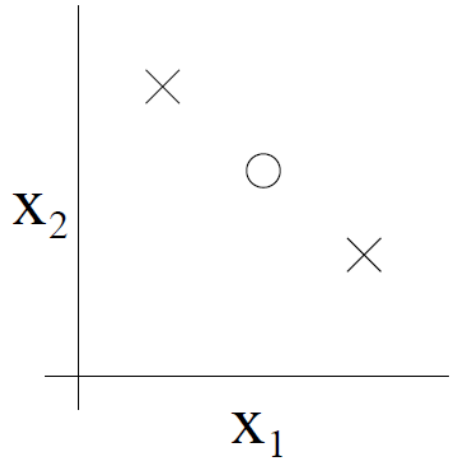
- Pode-se mostrar também que esta classe não pode *shatter* nenhum conjunto de 4 pontos.

- Portanto, o maior conjunto *shattered* por $\mathcal{H}$ tem tamanho 3 e, assim, $\mathrm{VC}(\mathcal{H}) = 3$.

- O fato de que $\mathrm{VC}(\mathcal{H}) = 3$ não implica que todo conjunto de tamanho 3 pode ser *shattered*.

- Veja o exemplo de pontos colineares abaixo, em que não há nenhum separador linear para esta rotulagem:



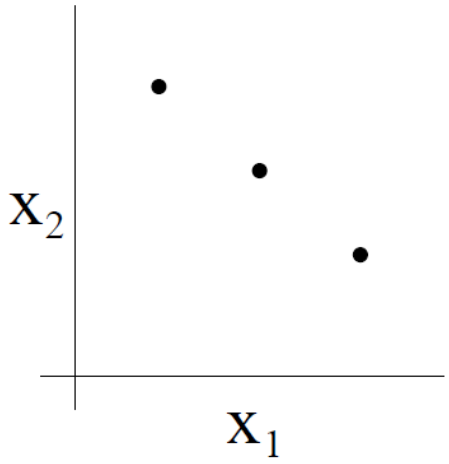


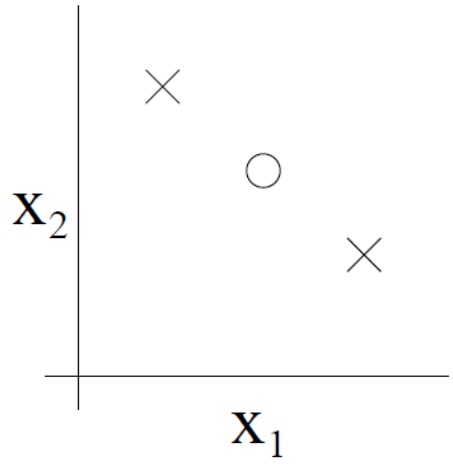
- Pode-se mostrar também que esta classe não pode *shatter* nenhum conjunto de 4 pontos.

- Portanto, o maior conjunto *shattered* por $\mathcal{H}$ tem tamanho 3 e, assim, $\mathrm{VC}(\mathcal{H}) = 3$.

- O fato de que $\mathrm{VC}(\mathcal{H}) = 3$ não implica que todo conjunto de tamanho 3 pode ser *shattered*.

- Veja o exemplo de pontos colineares abaixo, em que não há nenhum separador linear para esta rotulagem:



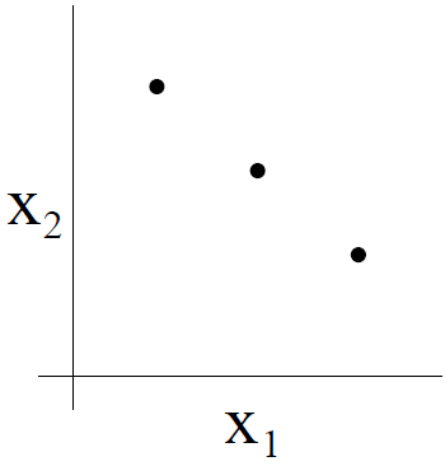


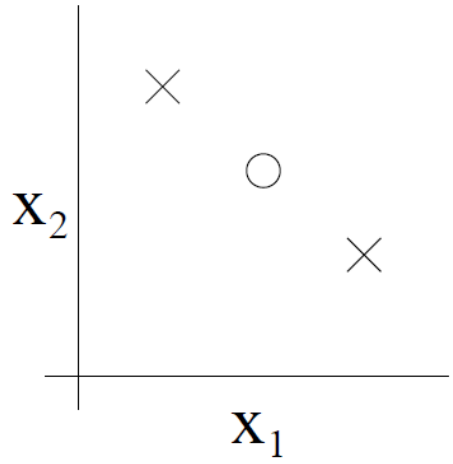
- Pode-se mostrar também que esta classe não pode *shatter* nenhum conjunto de 4 pontos.

- Portanto, o maior conjunto *shattered* por $\mathcal{H}$ tem tamanho 3 e, assim, $\mathrm{VC}(\mathcal{H}) = 3$.

- O fato de que $\mathrm{VC}(\mathcal{H}) = 3$ não implica que todo conjunto de tamanho 3 pode ser *shattered*.

- Veja o exemplo de pontos colineares abaixo, em que não há nenhum separador linear para esta rotulagem:



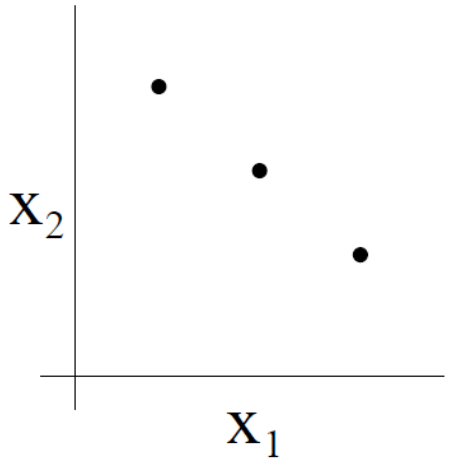


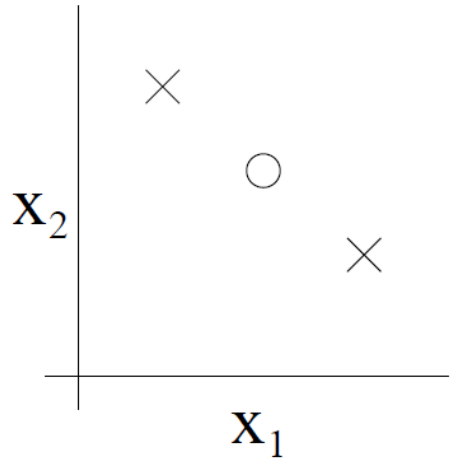
- Pode-se mostrar também que esta classe não pode *shatter* nenhum conjunto de 4 pontos.

- Portanto, o maior conjunto *shattered* por $\mathcal{H}$ tem tamanho 3 e, assim, $\mathrm{VC}(\mathcal{H}) = 3$.

- O fato de que $\mathrm{VC}(\mathcal{H}) = 3$ não implica que todo conjunto de tamanho 3 pode ser *shattered*.

- Veja o exemplo de pontos colineares abaixo, em que não há nenhum separador linear para esta rotulagem:





- Ou seja, para mostrar que $\mathrm{VC}(\mathcal{H})$ é ao menos **D**, precisamos mostrar que há no mínimo **um** conjunto de tamanho **D** que $\mathcal{H}$ pode *shatter*.

- Temos então o seguinte teorema, devido a Vapnik, e considerado por muitos como o mais importante de toda a teoria do aprendizado.

Teorema

Seja $\mathcal{H}$ dado e seja $\mathbf{D} = \mathrm{VC}(\mathcal{H})$. Então, com probabilidade no mínimo $1 - \delta$, temos para todo $h \in \mathcal{H}$:

$$|\varepsilon(h) - \hat{\varepsilon}(h)| \leq \mathcal{O}\left(\sqrt{\frac{\mathbf{D}}{n}\log\frac{n}{\mathbf{D}} + \frac{1}{n}\log\frac{1}{\delta}}\right).$$

E ainda, com probabilidade no mínimo $1 - \delta$, também temos

$$\varepsilon(\hat{h}) \leq \varepsilon(h^*) + \mathcal{O}\left(\sqrt{\frac{\mathbf{D}}{n}\log\frac{n}{\mathbf{D}} + \frac{1}{n}\log\frac{1}{\delta}}\right).$$

- Ou seja, para mostrar que $\mathrm{VC}(\mathcal{H})$ é ao menos **D**, precisamos mostrar que há no mínimo **um** conjunto de tamanho **D** que $\mathcal{H}$ pode *shatter*.

- Temos então o seguinte teorema, devido a Vapnik, e considerado por muitos como o mais importante de toda a teoria do aprendizado.

Teorema

Seja $\mathcal{H}$ dado e seja $\mathbf{D} = \mathrm{VC}(\mathcal{H})$. Então, com probabilidade no mínimo $1 - \delta$, temos para todo $h \in \mathcal{H}$:

$$|\varepsilon(h) - \hat{\varepsilon}(h)| \leq \mathcal{O}\left(\sqrt{\frac{\mathbf{D}}{n}\log\frac{n}{\mathbf{D}} + \frac{1}{n}\log\frac{1}{\delta}}\right).$$

E ainda, com probabilidade no mínimo $1 - \delta$, também temos

$$\varepsilon(\hat{h}) \leq \varepsilon(h^*) + \mathcal{O}\left(\sqrt{\frac{\mathbf{D}}{n}\log\frac{n}{\mathbf{D}} + \frac{1}{n}\log\frac{1}{\delta}}\right).$$

- Ou seja, para mostrar que $\mathrm{VC}(\mathcal{H})$ é ao menos $\mathbf{D}$, precisamos mostrar que há no mínimo **um** conjunto de tamanho $\mathbf{D}$ que $\mathcal{H}$ pode *shatter*.

- Temos então o seguinte teorema, devido a Vapnik, e considerado por muitos como o mais importante de toda a teoria do aprendizado.

Teorema

Seja $\mathcal{H}$ dado e seja $\mathbf{D} = \mathrm{VC}(\mathcal{H})$. Então, com probabilidade no mínimo $1 - \delta$, temos para todo $h \in \mathcal{H}$:

$$|\varepsilon(h) - \hat{\varepsilon}(h)| \leq \mathcal{O}\left(\sqrt{\frac{\mathbf{D}}{n}\log\frac{n}{\mathbf{D}} + \frac{1}{n}\log\frac{1}{\delta}}\right).$$

E ainda, com probabilidade no mínimo $1 - \delta$, também temos

$$\varepsilon(\hat{h}) \leq \varepsilon(h^*) + \mathcal{O}\left(\sqrt{\frac{\mathbf{D}}{n}\log\frac{n}{\mathbf{D}} + \frac{1}{n}\log\frac{1}{\delta}}\right).$$

- Ou seja, se a classe de hipóteses tem dimensão VC finita, então temos convergência uniforme para n grande.

- Como antes, isso nos permite estabelecer limitantes sobre $\varepsilon(h)$ em termos de $\varepsilon(h^*)$.

- Temos então o corolário seguinte:

Corolário

Para que $|\varepsilon(h) - \varepsilon(\hat{h})| \leq \gamma$ seja válido para todo $h \in \mathcal{H}$ (e portanto $\varepsilon(\hat{h}) \leq \varepsilon(h^*) + 2\gamma$) com probabilidade no mínimo $1 - \delta$, basta que $n = \mathcal{O}_{\gamma,\delta}(\mathbf{D})$.

- Ou seja, o número de exemplos de treino para aprender "bem" usando $\mathcal{H}$ é linear em $\mathrm{VC}(\mathcal{H})$.

- Ou seja, se a classe de hipóteses tem dimensão VC finita, então temos convergência uniforme para n grande.

- Como antes, isso nos permite estabelecer limitantes sobre $\varepsilon(h)$ em termos de $\varepsilon(h^*)$.

- Temos então o corolário seguinte:

Corolário

Para que $|\varepsilon(h) - \varepsilon(\hat{h})| \leq \gamma$ seja válido para todo $h \in \mathcal{H}$ (e portanto $\varepsilon(\hat{h}) \leq \varepsilon(h^*) + 2\gamma$) com probabilidade no mínimo $1 - \delta$, basta que $n = \mathcal{O}_{\gamma,\delta}(\mathbf{D})$.

- Ou seja, o número de exemplos de treino para aprender "bem" usando $\mathcal{H}$ é linear em $\mathrm{VC}(\mathcal{H})$.

- Ou seja, se a classe de hipóteses tem dimensão VC finita, então temos convergência uniforme para n grande.

- Como antes, isso nos permite estabelecer limitantes sobre $\varepsilon(h)$ em termos de $\varepsilon(h^*)$.

- Temos então o corolário seguinte:

Corolário

Para que $|\varepsilon(h) - \varepsilon(\hat{h})| \leq \gamma$ seja válido para todo $h \in \mathcal{H}$ (e portanto $\varepsilon(\hat{h}) \leq \varepsilon(h^*) + 2\gamma$) com probabilidade no mínimo $1 - \delta$, basta que $n = \mathcal{O}_{\gamma,\delta}(\mathbf{D})$.

- Ou seja, o número de exemplos de treino para aprender "bem" usando $\mathcal{H}$ é linear em $\mathrm{VC}(\mathcal{H})$.

- Ou seja, se a classe de hipóteses tem dimensão VC finita, então temos convergência uniforme para n grande.
- Como antes, isso nos permite estabelecer limitantes sobre $\varepsilon(h)$ em termos de $\varepsilon(h^*)$.
- Temos então o corolário seguinte:

Corolário

Para que $|\varepsilon(h) - \varepsilon(\hat{h})| \leq \gamma$ seja válido para todo $h \in \mathcal{H}$ (e portanto $\varepsilon(\hat{h}) \leq \varepsilon(h^*) + 2\gamma$) com probabilidade no mínimo $1 - \delta$, basta que $n = \mathcal{O}_{\gamma,\delta}(\mathbf{D})$.

- Ou seja, o número de exemplos de treino para aprender "bem" usando $\mathcal{H}$ é linear em $\mathrm{VC}(\mathcal{H})$.

- Ou seja, se a classe de hipóteses tem dimensão VC finita, então temos convergência uniforme para n grande.

- Como antes, isso nos permite estabelecer limitantes sobre $\varepsilon(h)$ em termos de $\varepsilon(h^*)$.

- Temos então o corolário seguinte:

Corolário

Para que $|\varepsilon(h) - \varepsilon(\hat{h})| \leq \gamma$ seja válido para todo $h \in \mathcal{H}$ (e portanto $\varepsilon(\hat{h}) \leq \varepsilon(h^*) + 2\gamma$) com probabilidade no mínimo $1 - \delta$, basta que $n = \mathcal{O}_{\gamma,\delta}(\mathbf{D})$.

- Ou seja, o número de exemplos de treino para aprender "bem" usando $\mathcal{H}$ é linear em $\mathrm{VC}(\mathcal{H})$.

- Ocorre que, para a maioria das classes de hipóteses, a dimensão VC é aproximadamente linear no número de parâmetros (assumindo-se uma "parametrização" razoável).

- Juntando as peças, concluímos que, para uma classe $\mathcal{H}$ (e um algoritmo que visa minimizar o erro de treino), o número de exemplos de treino para que o erro de generalização fique próximo ao do classificador ótimo é usualmente aproximadamente linear no número de parâmetros de $\mathcal{H}$.

- Ocorre que, para a maioria das classes de hipóteses, a dimensão VC é aproximadamente linear no número de parâmetros (assumindo-se uma "parametrização" razoável).

- Juntando as peças, concluímos que, para uma classe $\mathcal{H}$ (e um algoritmo que visa minimizar o erro de treino), o número de exemplos de treino para que o erro de generalização fique próximo ao do classificador ótimo é usualmente aproximadamente linear no número de parâmetros de $\mathcal{H}$.