

MLP

SVM

Universidade de São Paulo

Instituto de Ciências Matemáticas e de Computação

Departamento de Ciências de Computação

Rodrigo Fernandes de Mello

<http://www.icmc.usp.br/~mello>mello@icmc.usp.brID3
C4.5
J48

KNN

Algoritmos de Aprendizado Supervisionado

Algoritmos de Aprendizado Supervisionado

MLP

SVM

ID3
C4.5
J48

KNN

MLP

SVM

Mas como podemos provar que
esses algoritmos aprendem algo?ID3
C4.5
J48

KNN

Algoritmos de Aprendizado Supervisionado

Conceitos Básicos

MLP

SVM

Essa é a principal
Motivação para
A Teoria do
Aprendizado EstatísticoID3
C4.5
J48

KNN

Conceitos Básicos

Principais Fundadores: Vapnik e Chervonenkis

Conceitos Básicos

Principais Fundadores: Vapnik e Chervonenkis



Conceitos Básicos

Principais Fundadores: Vapnik e Chervonenkis



Como definiram aprendizado?

Conceitos Básicos

Principais Fundadores: Vapnik e Chervonenkis



Como definiram aprendizado?



Conceitos Básicos

Principais Fundadores: Vapnik e Chervonenkis



Como definiram aprendizado?



Processo de inferir regras gerais
a partir da observação de exemplos

Conceitos Básicos

Principais Fundadores: Vapnik e Chervonenkis



Como definiram aprendizado?



Processo de inferir regras gerais
a partir da observação de exemplos



Principais Fundadores: Vapnik e Chervonenkis



Como definiram aprendizado?



Processo de inferir regras gerais a partir da observação de exemplos



Isso leva a problemas de Classificação, ou seja, problemas de aprendizado supervisionado



Ninguém precisa definir formalmente um carro, a criança aprende por exemplos e rótulos

Quais features (características) podemos utilizar para classificar um objeto como um carro?

Quais features (características) podemos utilizar para classificar um objeto como um carro?

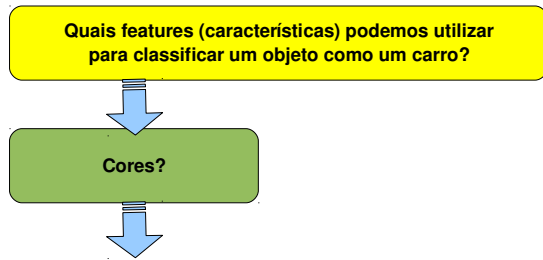


Quais features (características) podemos utilizar para classificar um objeto como um carro?

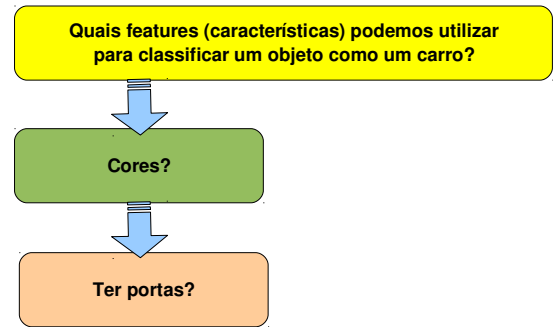


Cores?

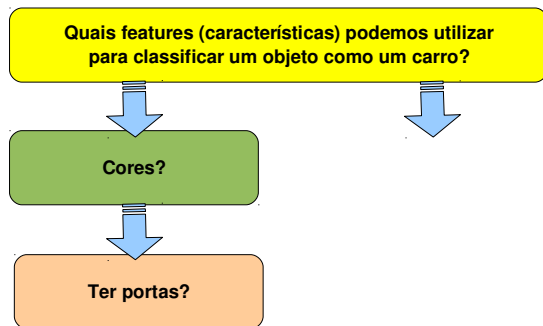
Conceitos Básicos



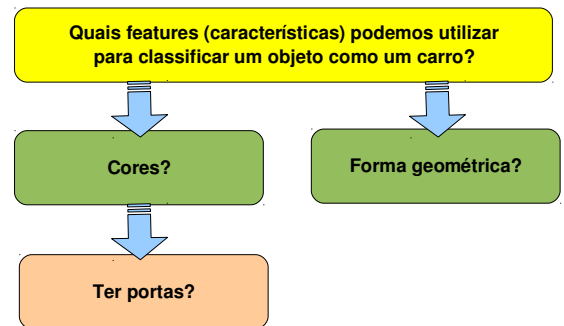
Conceitos Básicos



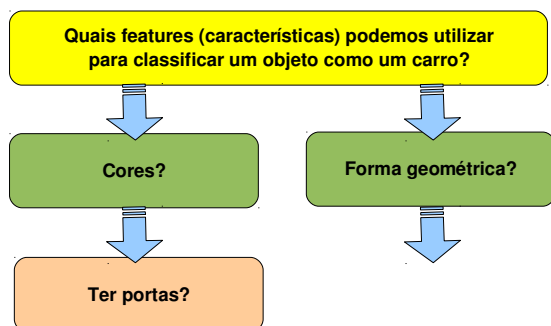
Conceitos Básicos



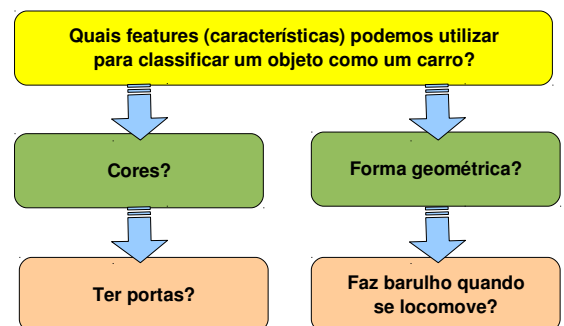
Conceitos Básicos

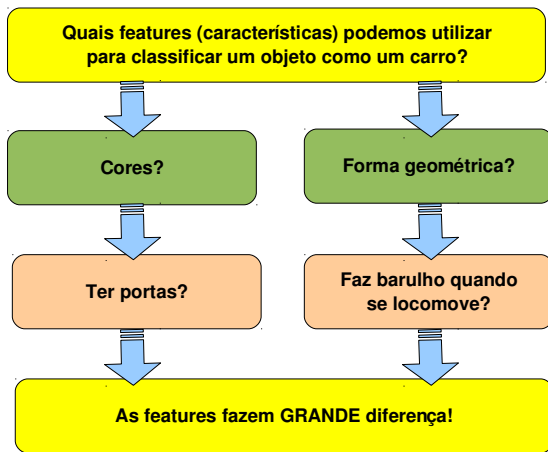


Conceitos Básicos



Conceitos Básicos

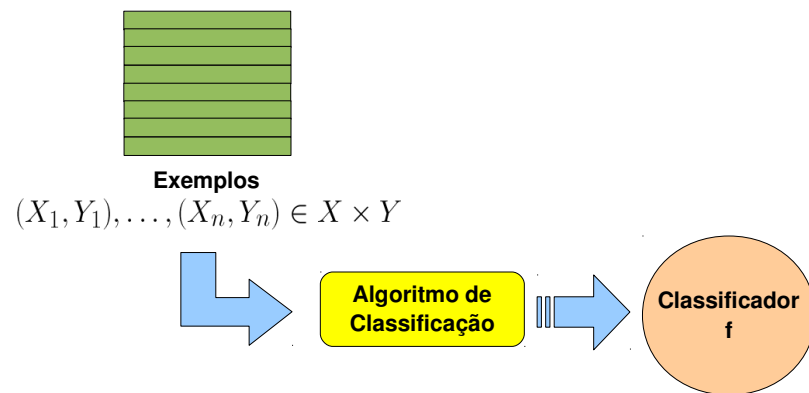




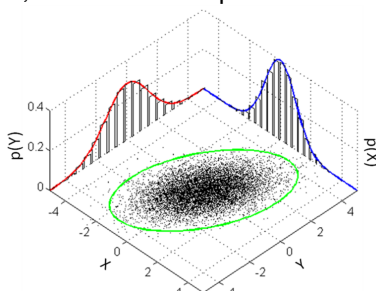
- Como etapa inicial, deve-se definir:
 - Espaço de entradas
 - Espaço de saídas
 - Considerações (Assumptions)
 - Função de Perda
 - Risco de um classificador

- Em **aprendizado supervisionado** temos:
 - O espaço de entradas X
 - O espaço de saídas ou rótulos Y
 - Considerando classificação binária, o espaço de rótulos é definido pelo conjunto $\{-1, +1\}$
- O aprendizado consiste, portanto, em **estimar**:

$$f : X \rightarrow Y$$
- Em que f é denominado **classificador**
 - Classificador é diferente de Algoritmo de Classificação



- Considera-se uma distribuição conjunta P sobre $X \times Y$
 - Para compreender essa distribuição conjunta considere duas variáveis aleatórias X e Y
 - Essa distribuição conjunta representa a probabilidade de cada X, Y estar em um particular intervalo



- Exemplo de distribuição conjunta P
 - Considere o lançamento de um dado
 - Seja $X=1$ se um número for PAR (i.e. 2, 4, or 6)
 - Seja $X=0$ se for ÍMPAR
 - Além disso, seja $Y=1$ se o número é PRIMO (i.e. 2, 3, or 5) e $Y=0$ caso contrário
 - Então, a probabilidade conjunta de X e Y é dada por:

- Exemplo de distribuição conjunta P
 - Considere o lançamento de um dado
 - Seja $X=1$ se um número for PAR (i.e. 2, 4, or 6)
 - Seja $X=0$ se for ÍMPAR
 - Além disso, seja $Y=1$ se o número é PRIMO (i.e. 2, 3, or 5) e $Y=0$ caso contrário
 - Então, a probabilidade conjunta de X e Y é dada por:

$$P(X=0, Y=0) = P\{1\} = \frac{1}{6} \quad P(X=1, Y=0) = P\{4, 6\} = \frac{2}{6}$$

$$P(X=0, Y=1) = P\{3, 5\} = \frac{2}{6} \quad P(X=1, Y=1) = P\{2\} = \frac{1}{6}$$

http://en.wikipedia.org/wiki/Joint_probability_distribution

http://en.wikipedia.org/wiki/Joint_probability_distribution

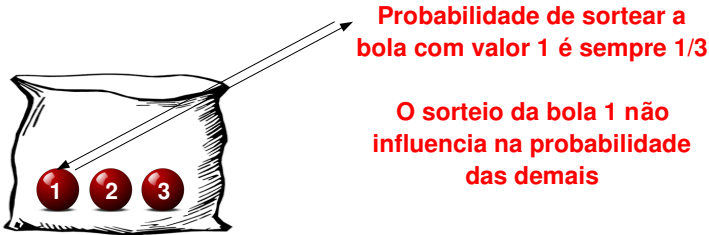
- Mas o que significa essa distribuição conjunta P?
 - Em Estatística: representa o relacionamento entre a variável aleatória X e a variável aleatória Y
 - Em Aprendizado de Máquina: representa o relacionamento entre os atributos de exemplos X e os rótulos Y

- Mas o que significa essa distribuição conjunta P?
 - Em Estatística: representa o relacionamento entre a variável aleatória X e a variável aleatória Y
 - Em Aprendizado de Máquina: representa o relacionamento entre os atributos de exemplos X e os rótulos Y

- Considera-se que os exemplos de treinamento são amostrados de maneira independente
 - Sendo assim, a obtenção de um primeiro exemplo com seu respectivo rótulo, i.e., (X_1, Y_1) , não influencia no sorteio seguinte

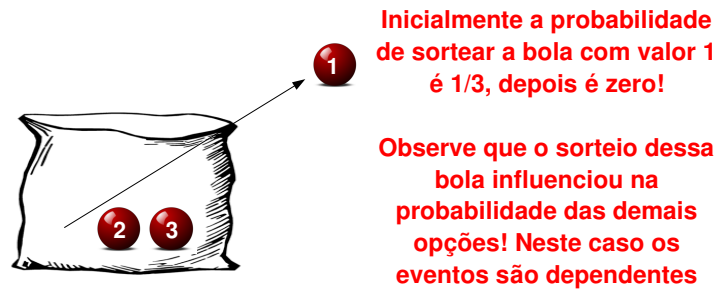
Primeiros passos da TAE

- Considera-se que os exemplos de treinamento são amostrados de maneira independente
 - Para ficar mais claro considere um sorteio com e sem reposição
- **SORTEIO COM REPOSIÇÃO**



Primeiros passos da TAE

- Considera-se que os exemplos de treinamento são amostrados de maneira independente
 - Para ficar mais claro considere um sorteio com e sem reposição
- **SORTEIO SEM REPOSIÇÃO**



Primeiros passos da TAE

- Exemplo em R utilizando função de autocorrelação para verificar a dependência entre eventos

Primeiros passos da TAE

- Vapnik realizou, ainda, um conjunto de considerações (assumptions) para formalizar a Teoria do Aprendizado Estatístico:
 - Nenhuma suposição é feita sobre P
 - Rótulos podem ser não determinísticos, devido a ruídos nos dados e sobreposição de classes
 - A amostragem de exemplos é independente
 - A distribuição P é fixa
 - A distribuição P é desconhecida no momento do aprendizado

Primeiros passos da TAE

- **Nenhuma suposição é feita sobre P**
 - A TAE não faz nenhuma suposição sobre a distribuição de probabilidades conjunta P , i.e., P pode ser qualquer distribuição
 - Logo a TAE trabalha de maneira agnóstica, o que é diferente da estatística tradicional em que geralmente supõe-se que P pertence a uma certa família de distribuições:
 - consequentemente o objetivo é reduzido à estimação dos parâmetros para tal distribuição

Primeiros passos da TAE

- **Rótulos podem ser não determinísticos, devido a ruídos e sobreposição de classes:**
 - Há duas razões para isso:
 - Primeira: Os dados podem estar sujeitos a ruídos nos rótulos, ou seja, os rótulos Y_i que temos no conjunto de treinamento podem estar errados
 - Essa é uma suposição importante, por exemplo, indivíduos rotulando emails como spam podem errar
 - Claro que esperamos que apenas uma pequena parte dos rótulos esteja de fato errada

- **Rótulos podem ser não determinísticos, devido a ruídos e sobreposição de classes:**

- Há duas razões para isso:
 - Segunda: Classes podem estar sobrepostas
 - Por exemplo, considere o problema de classificar pessoas como sexo masculino/feminino com base em suas alturas
 - Claro que uma pessoa com 1,80 m pode, em princípio, ser de qualquer um dos sexos, portanto não podemos associar um rótulo único Y para a entrada $X = 1,80$

- **Rótulos podem ser não determinísticos, devido a ruídos e sobreposição de classes:**

- Para o propósito de aprendizado, não importa a razão pela qual os rótulos são não determinísticos, na prática o que queremos encontrar são as probabilidades condicionais:

$$P(Y = 1|X = x)$$

$$P(Y = -1|X = x)$$

- No caso de um pequeno ruído nos rótulos, essas probabilidades condicionais são próximas de 0 ou de 1
- No caso de um grande ruído, essas probabilidades podem se aproximar de 0,5, dificultando o aprendizado

- **Rótulos podem ser não determinísticos, devido a ruídos e sobreposição de classes:**

- O mesmo problema ocorre na classificação de pessoas por gênero segundo suas alturas, isso ocorre pois há grande sobreposição entre classes
- Por exemplo, considere:

$$P(Y = \text{"masculino"}|X = 1.70) = 0.6$$
 - Na média erramos em 40% dos casos

- Assim, o aprendizado torna-se cada vez mais difícil quando a probabilidade condicional aproxima-se de 0,5, tornando inevitável que o classificador produza um grande número de erros

- **Amostragem independente**

- A TAE assume que as instâncias são amostradas de maneira independente
- Por exemplo, considere um conjunto de dados de treinamento utilizado para **aprendermos caracteres escritos a mão**:
 - É seguro assumirmos que tais caracteres formam uma amostra independente de toda a população de caracteres escritos a mão
- Em contrapartida, considere o cenário de **descoberta de novos medicamentos**:
 - Este é um campo em que pesquisadores procuram por novos compostos úteis para resolver problemas de saúde
 - É muito caro extrair características (ou atributos) desses compostos a fim de verificarmos suas propriedades

- **Amostragem independente**

- Em contrapartida, considere o cenário de descoberta de novos medicamentos:
 - Como resultado, somente alguns compostos são submetidos a experimentos em laboratório a fim de verificar suas propriedades
 - Esses compostos são cuidadosamente selecionados com base no conhecimento de pesquisadores
 - Nesse caso, não podemos assumir que os compostos são uma amostra independente de todos os compostos químicos existentes, **pois são manualmente selecionados com base em um processo não aleatório**

- **Amostragem independente**

- Há algumas áreas dentro de Aprendizado de Máquina que relaxam esse princípio de independência:
 - Ex: há pesquisadores que fazem **predição de séries temporais** considerando que observações são independentes entre si

- A distribuição **P** é fixa:

- A TAE **não assume o parâmetro “tempo”**, sendo assim não há alteração da distribuição **P** ao longo do tempo
- Esse cenário não é verificado, por exemplo, em:
 - Aprendizado de séries temporais
 - Mudança de conceito (concept drift) em fluxos de dados

- A distribuição **P** é desconhecida no momento do aprendizado

- Se conhecessemos **P**, o aprendizado seria trivial
 - Ele se reduziria a resolver uma regressão
- Assim, no aprendizado supervisionado temos acesso indireto a **P**
 - Isso significa que se temos um número suficientemente grande de exemplos de treinamento, podemos **estimar P**

- Conforme visto, o objetivo do **aprendizado supervisionado** é aprender uma função:

$$f : X \rightarrow Y$$

- Considerando:
 - Um espaço de entrada, instâncias ou objetos **X**
 - Um espaço de saída, de rótulos ou classes **Y**
- A essa função **f** dá-se o nome de **classificador**
- Para isso precisamos de uma medida sobre “**quão boa**” é essa função quando utilizada para classificar exemplos nunca vistos:
 - Assim introduzimos o conceito de **função de perda**

- Conforme visto, o objetivo do **aprendizado supervisionado** é aprender uma função:

$$f : X \rightarrow Y$$

- Considerando:
 - Um espaço de entrada, instâncias ou objetos **X**
 - Um espaço de saída, de rótulos ou classes **Y**
- A essa função **f** dá-se o nome de **classificador**
- Para isso precisamos de uma medida sobre “**quão boa**” é essa função quando utilizada para classificar exemplos nunca vistos:
 - Assim introduzimos o conceito de **função de perda**

Aqui surge um conceito importante e Essencial para a TAE!

Precisamos saber como um classificador Se comporta para exemplos nunca vistos!

$$l(x, y, f(x)) = \begin{cases} 1 & \text{se } f(x) \neq y \\ 0 & \text{caso contrário.} \end{cases}$$

- A **função de perda** mais simples é chamada de **0-1-loss** ou de **erro de classificação**:

- Assim a perda é 0 caso **x** seja classificado corretamente como **y**
- e essa perda é 1 caso contrário

$$l(x, y, f(x)) = \begin{cases} 1 & \text{se } f(x) \neq y \\ 0 & \text{caso contrário.} \end{cases}$$

- Há outros problemas, tais como a regressão de uma função, em que erros devem ser medidos na forma de números reais, assim, pode-se utilizar uma função de perda baseada em erros quadráticos:

$$l(x, y, f(x)) = (y - f(x))^2$$

- **Ou seja, a função de perda pode variar conforme o objetivo do aprendizado.** A única convenção é que zero indica classificação perfeita e valores maiores representam perdas no desempenho do classificador

Risco Esperado

- Enquanto a **função de perda** mede o erro de uma função f sobre um exemplo individual x , o **risco esperado** calcula a perda média de exemplos segundo a distribuição conjunta P :

$$R(f) = E(l(x, y, f(x)))$$

- Dessa maneira, uma **função f é melhor que outra g** caso:

$$R(f) < R(g)$$

- Assim, o melhor classificador f é dado pelo menor valor para $R(f)$, ou seja, aquele que apresenta menor **risco esperado**

Risco Esperado

- Dessa maneira, uma **função f é melhor que outra g** caso:

$$R(f) < R(g)$$

Observe que isso faz sentido!

Ou seja, o classificador com menor **Risco** segundo a distribuição conjunta P de exemplos versus rótulos

Risco Esperado

- Dessa maneira, uma **função f é melhor que outra g** caso:

$$R(f) < R(g)$$

O Problema é que não temos como calcular o **Risco Esperado** uma vez que assumimos desconhecer a distribuição conjunta P !!!

Risco Empírico

- Sabe-se que temos um **conjunto finito de exemplos** para obtermos um bom classificador f , ou seja, não temos todo universo de exemplos
 - Logo, não podemos calcular $R(f)$ para um dado classificador f
 - No entanto, podemos utilizar o conceito de **Risco Empírico** ou **Erro de Treinamento** na forma:

$$R_{emp}(f) = \frac{1}{n} \sum_{i=1}^n l(x_i, y_i, f(x_i))$$

Risco Empírico

- Sabe-se que temos um **conjunto finito de exemplos** para obtermos um bom classificador f , ou seja, não temos todo universo de exemplos

Aqui começa, de fato, o **Princípio de Minimização do Risco Empírico**.

Esse Princípio visa escolher o melhor classificador com base no **Risco Empírico**!

$$f_n := \operatorname{argmin}_{f \in \mathcal{F}} R_{emp}(f)$$

Esse conceito é base para a TAE!

- Esperamos que um bom classificador f produza um baixo **Risco Empírico** para os exemplos de treinamento
 - Caso contrário o classificador não é nem capaz de representar os exemplos de treinamento
 - Mas será que minimizando o Risco Empírico obtemos um bom classificador?
 - Um classificador pode ser muito específico ou especializado no conjunto de treinamento e produzir um **Risco Empírico** mínimo, no entanto, seu **Risco Esperado** seria muito grande

- Esperamos que um bom classificador f produza um baixo **Risco Empírico** para os exemplos de treinamento
 - Caso contrário o classificador não é nem capaz de representar os exemplos de treinamento
 - Mas será que minimizando o Risco Empírico obtemos um bom classificador?
 - Um classificador pode ser muito específico ou especializado no conjunto de treinamento e produzir um **Risco Empírico** mínimo, no entanto, seu **Risco Esperado** seria muito grande

- Esperamos que um bom classificador f produza um baixo **Risco Empírico** para os exemplos de treinamento
 - Caso contrário o classificador não é nem capaz de representar os exemplos de treinamento
 - Mas será que minimizando o Risco Empírico obtemos um bom classificador?
 - Um classificador pode ser muito específico ou especializado no conjunto de treinamento e produzir um **Risco Empírico** mínimo, no entanto, seu **Risco Esperado** seria muito grande

- Esperamos que um bom classificador f produza um baixo **Risco Empírico** para os exemplos de treinamento
 - Caso contrário o classificador não é nem capaz de representar os exemplos de treinamento
 - Mas será que minimizando o Risco Empírico obtemos um bom classificador?
 - Um classificador pode ser muito específico ou especializado no conjunto de treinamento e produzir um **Risco Empírico** mínimo, no entanto, seu **Risco Esperado** seria muito grande

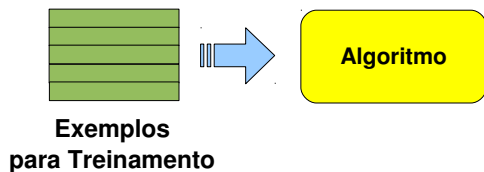
- Classificador baseado em memória

Isso torna o Princípio de Minimização do Risco Empírico Inconsistente!

Vejamos mais usando um exemplo...

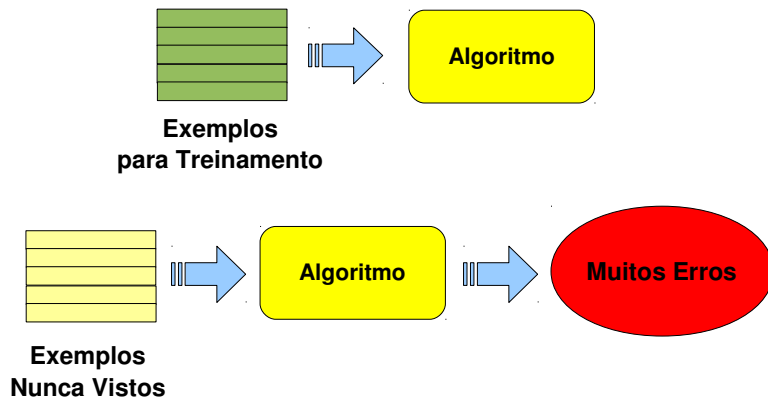
Risco Empírico

- Classificador baseado em memória



Risco Empírico

- Classificador baseado em memória

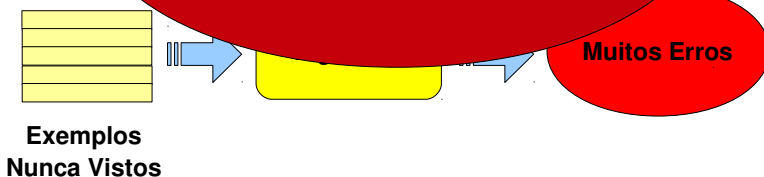


Risco Empírico

- Classificador baseado em memória

Sendo assim o Risco Empírico não é um bom estimador para o Risco Esperado!!

Portanto o PMRE é inconsistente, o que é observado por este exemplo bem simples!



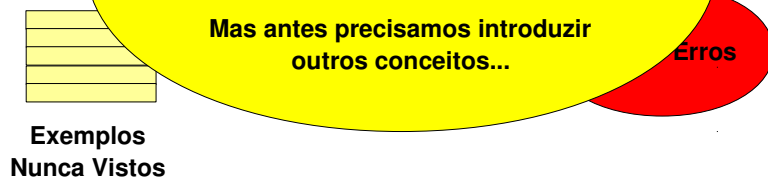
Risco Empírico

- Classificador baseado em memória

Mas há como resgatar esse Princípio e torná-lo válido para certos cenários?

Veremos...

Mas antes precisamos introduzir outros conceitos...



Generalização

- Assim surge o conceito de **Generalização** de um classificador f como:

$$|R(f_n) - R_{emp}(f_n)|$$

- Dessa maneira, um classificador generaliza bem quando essa diferença é pequena, ou seja, **o Risco Empírico é próximo ao Risco Esperado**
- Um classificador com boa generalização não necessariamente produz um baixo valor para Risco Empírico nem mesmo para o Risco Esperado

Generalização

- Assim surge o conceito de **Generalização** de um classificador f como:

$$|R(f_n) - R_{emp}(f_n)|$$

- Sendo assim, um Classificador com Boa Generalização é aquele em que seu Risco Empírico é um bom estimador para o Risco Esperado
- Um classificador com boa generalização não necessariamente produz um baixo valor para Risco Empírico nem mesmo para o Risco Esperado

Dilema Bias-Variância

- Por exemplo, considere os exemplos (pontos) coletados em um experimento qualquer e duas funções distintas para representar tais exemplos (uma reta e uma função polinomial de ordem maior que 1)

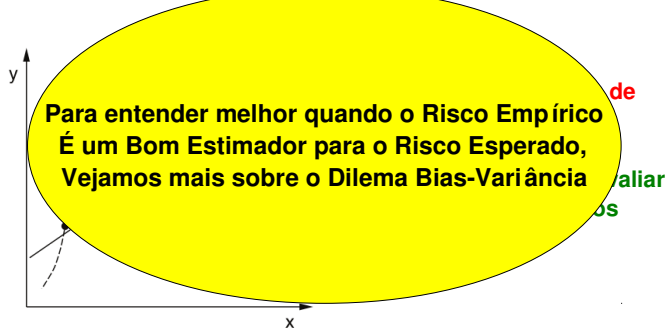
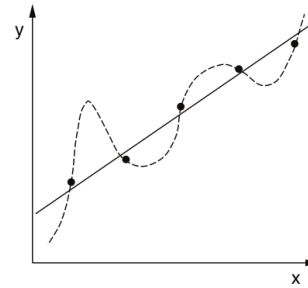


Figura obtida de Luxburg and Scholkopf, Statistical Learning Theory: Models, Concepts, and Results

Dilema Bias-Variância

- Por exemplo, considere os exemplos (pontos) coletados em um experimento qualquer e duas funções distintas para representar tais exemplos (uma reta e uma função polinomial de ordem maior que 1)



Qual das duas funções de regressão é a melhor?

Para isso precisamos avaliar seus respectivos Riscos Esperados

Figura obtida de Luxburg and Scholkopf, Statistical Learning Theory: Models, Concepts, and Results

Dilema Bias-Variância

- No entanto, como computar o Risco Esperado se somente temos uma amostra?

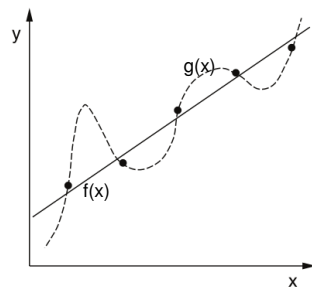


Figura obtida de Luxburg and Scholkopf, Statistical Learning Theory: Models, Concepts, and Results

Dilema Bias-Variância

- No entanto, como computar o Risco Esperado se somente temos uma amostra?

• Considere que a reta $f(x)$ tem Risco Empírico maior que zero

• Considere que a função $g(x)$ tem Risco Empírico igual a zero

• Considere que $g(x)$ representa um modelo com overfitting

• Logo, conforme exemplos nunca vistos são recebidos, $g(x)$ sai de um Risco Empírico igual a zero para um Risco Esperado muito alto

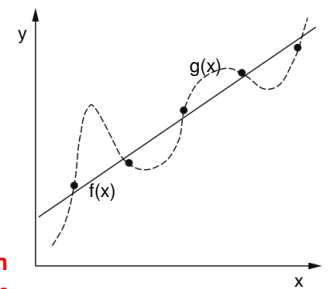


Figura obtida de Luxburg and Scholkopf, Statistical Learning Theory: Models, Concepts, and Results

Dilema Bias-Variância

- No entanto, como computar o Risco Esperado se somente temos uma amostra?

• Considere que a reta $f(x)$ tem Risco Empírico maior que zero

• Considere que a função $g(x)$ tem Risco Empírico igual a zero

• Considere que $g(x)$ representa um modelo com overfitting

• Logo, conforme exemplos nunca vistos são recebidos, $g(x)$ sai de um Risco Empírico igual a zero para um Risco Esperado muito alto

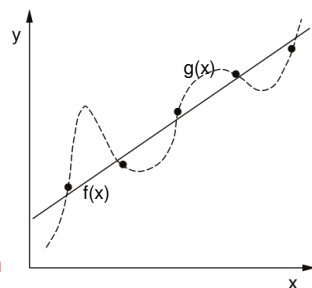


Figura obtida de Luxburg and Scholkopf, Statistical Learning Theory: Models, Concepts, and Results

Dilema Bias-Variância

- No entanto, como computar o Risco Esperado se somente temos uma amostra?

• Considere que a reta $f(x)$ tem Risco Empírico maior que zero

• Considere que a função $g(x)$ tem Risco Empírico igual a zero

• Considere que $g(x)$ representa um modelo com overfitting

• Logo, conforme exemplos nunca vistos são recebidos, $g(x)$ sai de um Risco Empírico igual a zero para um Risco Esperado muito alto

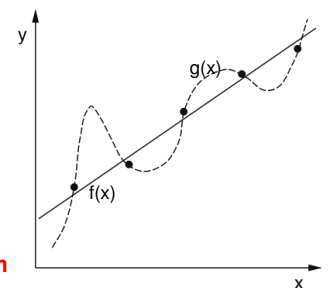


Figura obtida de Luxburg and Scholkopf, Statistical Learning Theory: Models, Concepts, and Results

Dilema Bias-Variância

- No entanto, como computar o Risco Esperado se somente temos uma amostra?
- Considere que a reta $f(x)$ tem Risco Empírico maior que zero
- Considere que a função $g(x)$ tem Risco Empírico igual a zero
- Considere que $g(x)$ representa um modelo com overfitting
- Logo, conforme exemplos nunca vistos são recebidos, $g(x)$ sai de um Risco Empírico igual a zero para um Risco Esperado muito alto

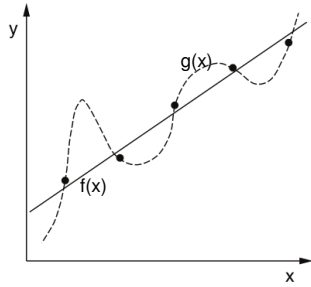


Figura obtida de Luxburg and Scholkopf, Statistical Learning Theory: Models, Concepts, and Results

Dilema Bias-Variância

- No entanto, como computar o Risco Esperado se somente temos uma amostra?
- Considere que a reta $f(x)$ tem Risco Empírico maior que zero
- Considere que a função $g(x)$ tem Risco Empírico igual a zero
- Considere que $g(x)$ representa um modelo com overfitting
- Logo, conforme exemplos nunca vistos são recebidos, $g(x)$ sai de um Risco Empírico igual a zero para um Risco Esperado muito alto

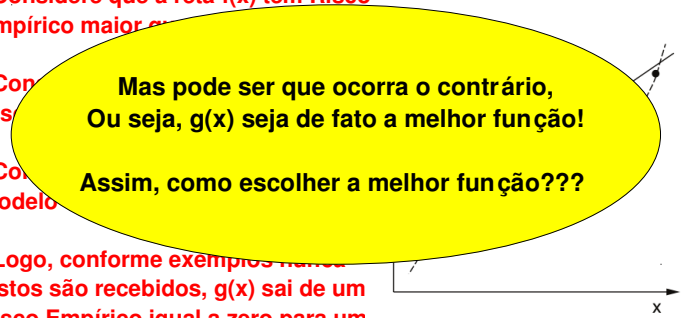


Figura obtida de Luxburg and Scholkopf, Statistical Learning Theory: Models, Concepts, and Results

Dilema Bias-Variância

- Finalmente, qual classificador devemos escolher?

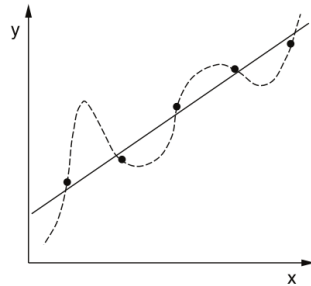


Figura obtida de Luxburg and Scholkopf, Statistical Learning Theory: Models, Concepts, and Results

Dilema Bias-Variância

- Finalmente, qual classificador devemos escolher?

1) Aquele que tem maior aptidão com os dados de treinamento (mais complexo)?

2) ou aquele que apresenta maior Risco Empírico porém é obtido por uma função mais simples?

Em Estatística esse problema é denominado Dilema Bias-Variância

Esse Dilema é abordado pela TAE!

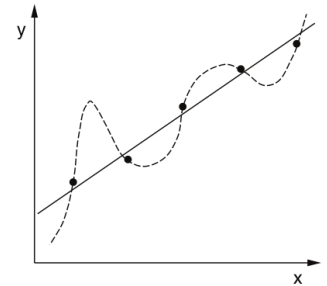


Figura obtida de Luxburg and Scholkopf, Statistical Learning Theory: Models, Concepts, and Results

Dilema Bias-Variância

- Finalmente, qual classificador devemos escolher?

1) Aquele que tem maior aptidão com os dados de treinamento (mais complexo)?

2) ou aquele que apresenta maior Risco Empírico porém é obtido por uma função mais simples?

Em Estatística esse problema é denominado Dilema Bias-Variância

Esse Dilema é abordado pela TAE!

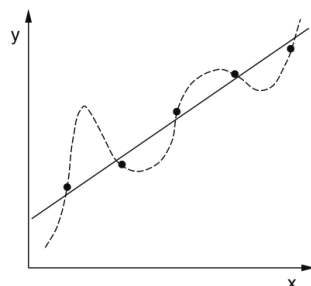


Figura obtida de Luxburg and Scholkopf, Statistical Learning Theory: Models, Concepts, and Results

Dilema Bias-Variância

- Finalmente, qual classificador devemos escolher?

1) Aquele que tem maior aptidão com os dados de treinamento (mais complexo)?

2) ou aquele que apresenta maior Risco Empírico porém é obtido por uma função mais simples?

Em Estatística esse problema é denominado Dilema Bias-Variância

Esse Dilema é abordado pela TAE!

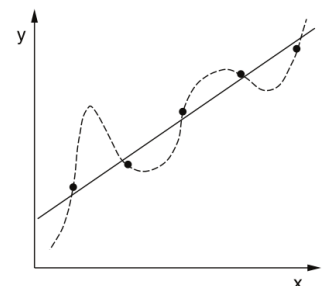


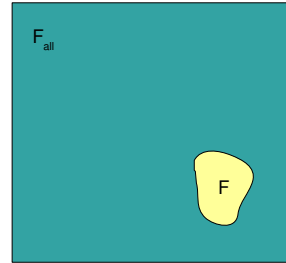
Figura obtida de Luxburg and Scholkopf, Statistical Learning Theory: Models, Concepts, and Results

Dilema Bias-Variância

- Dilema:
 - Bias:
 - Se assumimos um fit linear dos dados, apenas um classificador linear poderá ser obtido
 - Ou seja, há um forte viés (modelo linear) imposto por nós mesmos
 - Variância:
 - Se aproximamos um polinômio de ordem n dos dados de treinamento, sempre podemos chegar a um classificador perfeito para os dados da amostra
 - No entanto, esse classificador está sujeito a maiores flutuações para dados nunca vistos

Dilema Bias-Variância

- Uma dicotomia associada ao dilema Bias-Variância surge quando consideramos o espaço de possíveis funções para definir nosso classificador



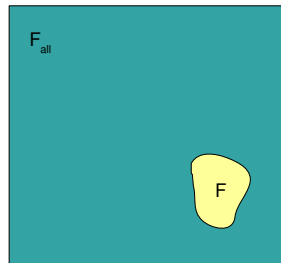
Dilema Bias-Variância

- Uma dicotomia associada ao dilema Bias-Variância surge quando consideramos o espaço de possíveis funções para definir nosso classificador

Considere o espaço F_{all} com todas as possíveis funções de classificação ou regressão

Poderíamos definir um forte viés, ou seja, um subespaço F contendo apenas as funções lineares para realizar a regressão

Esse aumento no viés, reduz a variância, i.e, reduz o número de possíveis classificadores que podemos obter



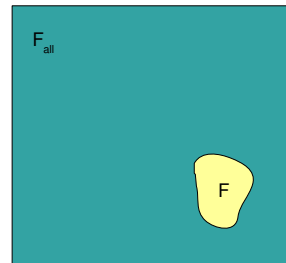
Dilema Bias-Variância

- Uma dicotomia associada ao dilema Bias-Variância surge quando consideramos o espaço de possíveis funções para definir nosso classificador

Considere o espaço F_{all} com todas as possíveis funções de classificação ou regressão

Poderíamos definir um forte viés, ou seja, um subespaço F contendo apenas as funções lineares para realizar a regressão

Esse aumento no viés, reduz a variância, i.e, reduz o número de possíveis classificadores que podemos obter



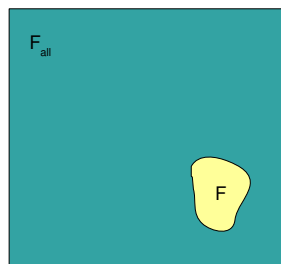
Dilema Bias-Variância

- Uma dicotomia associada ao dilema Bias-Variância surge quando consideramos o espaço de possíveis funções para definir nosso classificador

Considere o espaço F_{all} com todas as possíveis funções de classificação ou regressão

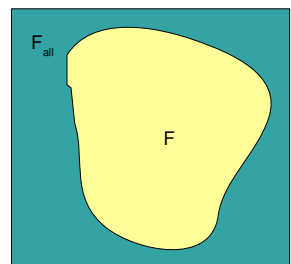
Poderíamos definir um forte viés, ou seja, um subespaço F contendo apenas as funções lineares para realizar a regressão

Esse aumento no viés, reduz a variância, i.e, reduz o número de possíveis classificadores que podemos obter



Dilema Bias-Variância

- Uma dicotomia associada ao dilema Bias-Variância surge quando consideramos o espaço de possíveis funções para definir nosso classificador

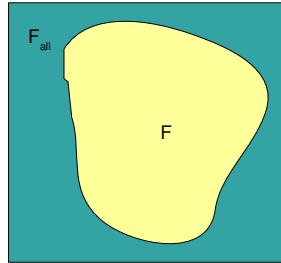


Dilema Bias-Variância

- Uma dicotomia associada ao dilema Bias-Variância surge quando consideramos o espaço de possíveis funções para definir nosso classificador

Em contrapartida, se o viés for pequeno, ou seja, o subespaço F contém muitas funções a serem consideradas

Nesse caso a variância é muito grande, i.e, o número de possíveis classificadores que podemos obter para um problema é muito grande!

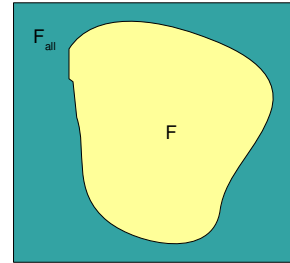


Dilema Bias-Variância

- Uma dicotomia associada ao dilema Bias-Variância surge quando consideramos o espaço de possíveis funções para definir nosso classificador

Em contrapartida, se o viés for pequeno, ou seja, o subespaço F contém muitas funções a serem consideradas

Nesse caso a variância é muito grande, i.e, o número de possíveis classificadores que podemos obter para um problema é muito grande!



Consistência

$$|R(f_n) - R_{emp}(f_n)|$$

Consistência

- Vapnik relaciona o conceito de Generalização com o de Consistência

- Sabe-se que o conceito de **Generalização** é definido por:

$$|R(f_n) - R_{emp}(f_n)|$$

- O conceito de **Consistência** não é particular de uma função f , mas sim de um conjunto de funções
 - Assim como na estatística, a noção de consistência visa realizar uma afirmação sobre o que ocorre quando número de exemplos tende ao infinito
- Isso significa que um algoritmo de aprendizado deveria **convergir** para o melhor classificador possível conforme o número de exemplos de treinamento aumenta

Consistência

- Vapnik relaciona o conceito de Generalização com o de Consistência

- Sabe-se que o conceito de **Generalização** é definido por:

$$|R(f_n) - R_{emp}(f_n)|$$

- O conceito de **Consistência** não é particular de uma função f , mas sim de um conjunto de funções
 - Assim como na estatística, a noção de consistência visa realizar uma afirmação sobre o que ocorre quando número de exemplos tende ao infinito
- Isso significa que um algoritmo de aprendizado deveria **convergir** para o melhor classificador possível conforme o número de exemplos de treinamento aumenta

Consistência

- Vapnik relaciona o conceito de Generalização com o de Consistência

- Sabe-se que o conceito de **Generalização** é definido por:

$$|R(f_n) - R_{emp}(f_n)|$$

- O conceito de **Consistência** não é particular de uma função f , mas sim de um conjunto de funções
 - Assim como na estatística, a noção de consistência visa realizar uma afirmação sobre o que ocorre quando número de exemplos tende ao infinito
- Isso significa que um algoritmo de aprendizado deveria **convergir** para o melhor classificador possível conforme o número de exemplos de treinamento aumenta

Consistência

- Vapnik relaciona o conceito de Generalização com o de Consistência
- Sabemos que, para um algoritmo de aprendizado ser consistente, ele deve satisfazer a seguinte condição:
 - Assim que o tamanho da amostra tende ao infinito, a consistência visa realizar uma classificação que se aproxime do melhor classificador possível.
- Isso significa que um algoritmo de aprendizado deveria **convergir** para o melhor classificador possível conforme o número de exemplos de treinamento aumenta

Depois da Generalização, Vapnik necessitou de Conceitos que permitissem provar que um classificador tende ao melhor dentro de um subespaço, conforme aumenta o número de exemplos de treinamento

Consistência

- Vapnik relaciona o conceito de Generalização com o de Consistência
- Sabemos que, para um algoritmo de aprendizado ser consistente, ele deve satisfazer a seguinte condição:
 - Assim que o tamanho da amostra tende ao infinito, a consistência visa realizar uma classificação que se aproxime do melhor classificador possível.
- Isso significa que um algoritmo de aprendizado deveria **convergir** para o melhor classificador possível conforme o número de exemplos de treinamento aumenta

Logo, há dois pontos importantes considerados por Vapnik:

- Encontrar uma forma que garanta que o Risco Empírico se aproxime do Risco Esperado
- Um Algoritmo de Aprendizado deve buscar pelo melhor classificador dentro de seu viés conforme o número de exemplos de treinamento aumenta

Consistência

Consistência

- Há, no entanto, dois **tipos de Consistência na literatura**:
 - **Consistência com respeito ao subespaço F**
 - Um algoritmo de aprendizado define um viés, ou seja, um subespaço de funções que considera para obter um classificador f
 - Um algoritmo é consistente com respeito ao subespaço F quando converge para o melhor classificador dentro desse subespaço conforme o tamanho da amostra aumenta
 - **Consistência de Bayes**
 - Dentro do espaço F_{all} há um classificador ideal (também denominado f_{Bayes} ou classificador de Bayes), ou seja, o melhor de todos
 - Um algoritmo é consistente com respeito a Bayes quando converge para o melhor classificador possível conforme o tamanho da amostra aumenta

Consistência

Consistência

- Há, no entanto, dois **tipos de Consistência na literatura**:
 - **Consistência com respeito ao subespaço F**
 - Um algoritmo de aprendizado define um viés, ou seja, um subespaço de funções que considera para obter um classificador f
 - Um algoritmo é consistente com respeito ao subespaço F quando converge para o melhor classificador dentro desse subespaço conforme o tamanho da amostra aumenta
 - **Consistência de Bayes**
 - Dentro do espaço F_{all} há um classificador ideal (também denominado f_{Bayes} ou classificador de Bayes), ou seja, o melhor de todos
 - Um algoritmo é consistente com respeito a Bayes quando converge para o melhor classificador possível conforme o tamanho da amostra aumenta

- Há, no entanto, dois tipos de Consistência na literatura:
 - **Consistência com respeito ao subespaço F**
 - O melhor classificador no subespaço F é dado por:

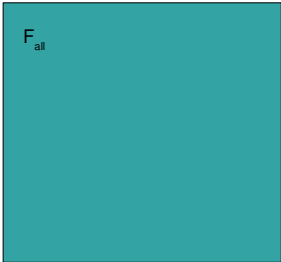
$$f_F = \arg \min_{f \in F} R(f)$$

- **Consistência de Bayes**
 - O classificador de Bayes é dado por:

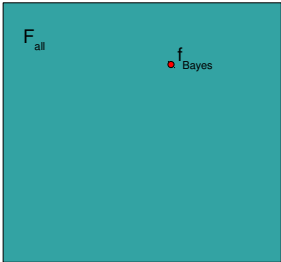
$$f_{Bayes} = \arg \min_{f \in F_{all}} R(f)$$

Consistência	Consistência
--------------	--------------

- Consistência com respeito ao subespaço F :

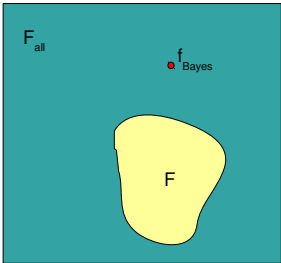


- Consistência com respeito ao subespaço F :

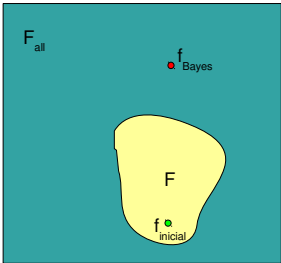


Consistência	Consistência
--------------	--------------

- Consistência com respeito ao subespaço F :

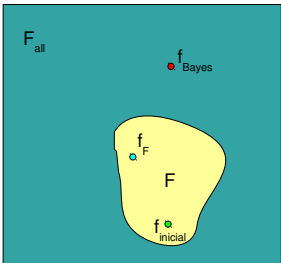


- Consistência com respeito ao subespaço F :

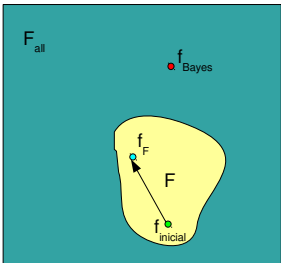


Consistência	Consistência
--------------	--------------

- Consistência com respeito ao subespaço F :



- Consistência com respeito ao subespaço F :



- Dizemos que um algoritmo de aprendizado é **consistente com respeito a F para uma certa distribuição de probabilidades P**:

- Se o Risco Esperado de um classificador f_n , isso é, $R(f_n)$, converge, em probabilidade, para o Risco Esperado $R(f_F)$ do melhor classificador em F, conforme o número de exemplos tende ao infinito:

$$P(R(f_n) - R(f_F) > \varepsilon) \rightarrow 0 \text{ conforme } n \rightarrow \infty$$

- Um algoritmo de aprendizado é dito ser **consistente com respeito a Bayes para uma certa distribuição de probabilidades P** se o Risco Esperado $R(f_n)$ para uma f_n converge, em probabilidade, para o Risco Esperado $R(f_{\text{Bayes}})$ do classificador de Bayes conforme o número de exemplos tende ao infinito

$$P(R(f_n) - R(f_{\text{Bayes}}) > \varepsilon) \rightarrow 0 \text{ conforme } n \rightarrow \infty$$

- Dizemos que um algoritmo de aprendizado é **universalmente consistente com respeito a F**:

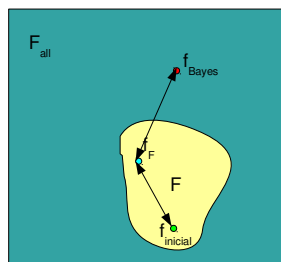
- Se o Risco Esperado de um classificador f_n , isso é, $R(f_n)$, converge para o Risco Esperado $R(f_F)$ do melhor classificador em F, conforme o número de exemplos tende ao infinito **para toda distribuição P**

- Este conceito de consistência é mais geral:**

- Não requer que consideremos uma certa distribuição de probabilidades P**
- Esta consistência é a que interessa para a Teoria do Aprendizado Estatístico**
- Lembre-se que assumimos não conhecer P!!!**

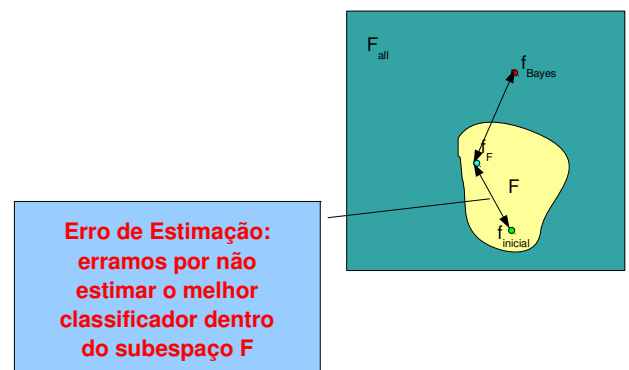
Paralelo entre Consistências e Erros de Aprendizado

- Paralelo entre Consistências e Erros



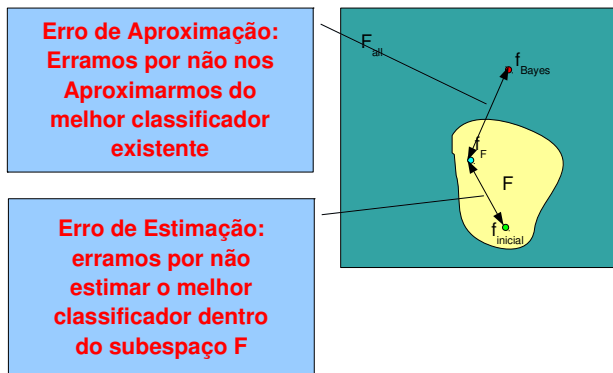
Paralelo entre Consistências e Erros de Aprendizado

- Paralelo entre Consistências e Erros



Paralelo entre Consistências e Erros de Aprendizado

- Paralelo entre Consistências e Erros



Paralelo entre Consistências e Erros de Aprendizado

- O Erro total pode ser definido em termos do Erro de Aproximação e do Erro de Estimação

$$R(f_n) - R(f_{Bayes}) = \underbrace{(R(f_n) - R(f_F))}_{\text{Erro de Estimação}} + \underbrace{(R(f_F) - R(f_{Bayes}))}_{\text{Erro de Aproximação}}$$

Paralelo entre Consistências e Erros de Aprendizado

Paralelo entre Consistências e Erros de Aprendizado

- Logo:
 - Se escolhermos um subespaço F muito grande, o termo de aproximação se tornará pequeno, no entanto, o termo de estimação será muito grande
 - O erro de estimação torna-se muito grande, pois funções mais complexas e, portanto, que geram **overfitting** aos dados, estarão nesse conjunto F
 - O oposto ocorre caso o espaço F seja muito pequeno

Paralelo entre Consistências e Erros de Aprendizado

Paralelo entre Consistências e Erros de Aprendizado

- Logo:
 - Se escolhermos um subespaço F muito grande, o termo de aproximação se tornará pequeno, no entanto, o termo de estimação será muito grande
 - O erro de estimação torna-se muito grande, pois funções mais complexas e, portanto, que geram **overfitting** aos dados, estarão nesse conjunto F
 - O oposto ocorre caso o espaço F seja muito pequeno

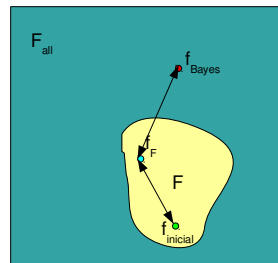
- Logo:
 - Se escolhermos um subespaço F muito grande, o termo de aproximação se tornará pequeno, no entanto, o termo de estimação será muito grande
 - O erro de estimação torna-se muito grande, pois funções mais complexas e, portanto, que geram **overfitting** aos dados, estarão nesse conjunto F
 - O oposto ocorre caso o espaço F seja muito pequeno

$$R(f_n) - R(f_{Bayes}) = \underbrace{(R(f_n) - R(f_F))}_{\text{Erro de Estimação}} + \underbrace{(R(f_F) - R(f_{Bayes}))}_{\text{Erro de Aproximação}}$$

Paralelo entre Consistências e Erros de Aprendizado

- Erro de estimação:
 - Resultado da incerteza existente nos dados de treinamento
 - Esse erro é também chamado de **variância** na estatística
- Erro de aproximação:
 - Resultado do viés do algoritmo de aprendizado
 - Na estatística, esse erro é também denominado **viés** (do inglês **bias**)

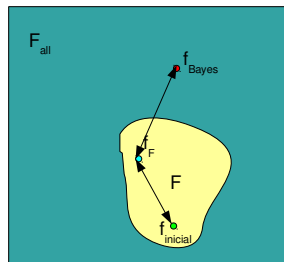
Paralelo entre Consistências e Erros de Aprendizado



$$R(f_n) - R(f_{Bayes}) = \underbrace{(R(f_n) - R(f_F))}_{\text{Erro de Estimação}} + \underbrace{(R(f_F) - R(f_{Bayes}))}_{\text{Erro de Aproximação}}$$

Paralelo entre Consistências e Erros de Aprendizado

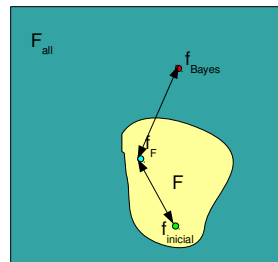
- Agora podemos notar que ao definir um subespaço F, definimos um equilíbrio entre os erros de estimação e de aproximação:
 - Um grande viés produz baixa variância ou pequeno erro de estimação, no entanto leva a um grande erro de aproximação
 - Um pequeno viés produz grande variância ou seja alto erro de estimação, no entanto leva a um pequeno erro de aproximação



$$R(f_n) - R(f_{Bayes}) = \underbrace{(R(f_n) - R(f_F))}_{\text{Erro de Estimação}} + \underbrace{(R(f_F) - R(f_{Bayes}))}_{\text{Erro de Aproximação}}$$

Paralelo entre Consistências e Erros de Aprendizado

- Agora podemos notar que ao definir um subespaço F, definimos um equilíbrio entre os erros de estimação e de aproximação:
 - Um grande viés produz baixa variância ou pequeno erro de estimação, no entanto leva a um grande erro de aproximação
 - Um pequeno viés produz grande variância ou seja alto erro de estimação, no entanto leva a um pequeno erro de aproximação



$$R(f_n) - R(f_{Bayes}) = \underbrace{(R(f_n) - R(f_F))}_{\text{Erro de Estimação}} + \underbrace{(R(f_F) - R(f_{Bayes}))}_{\text{Erro de Aproximação}}$$

Underfitting e Overfitting

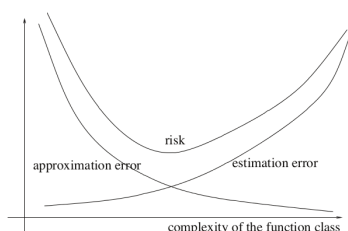


Figura obtida de Luxburg and Scholkopf, Statistical Learning Theory: Models, Concepts, and Results

Underfitting e Overfitting

- A partir dessa relação podemos definir:
 - **Underfitting** – Se o subespaço F é pequeno, o erro de estimação é menor, mas o erro de aproximação é muito grande
 - **Overfitting** – Se o subespaço F é grande, então o erro de estimação é muito grande, enquanto o erro de aproximação é pequeno
- O melhor para aprendizado de máquina é dado no ponto de equilíbrio
- Observe que a figura abaixo define o tamanho do subespaço F em termos da complexidade das funções contidas em F

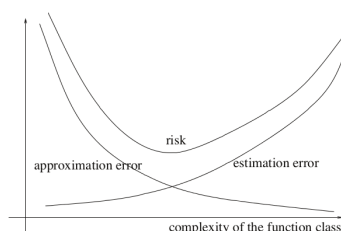


Figura obtida de Luxburg and Scholkopf, Statistical Learning Theory: Models, Concepts, and Results

Underfitting e Overfitting

• A partir dessa relação podemos definir:

- **Underfitting** – Se o subespaço F é pequeno, o erro de estimação é menor, mas o erro de aproximação é muito grande
- **Overfitting** – Se o subespaço F é grande, então o erro de estimação é muito grande, enquanto o erro de aproximação é pequeno
- O melhor para aprendizado de máquina é dado no ponto de equilíbrio
- Observe que a figura abaixo define o tamanho do subespaço F em termos da complexidade das funções contidas em F

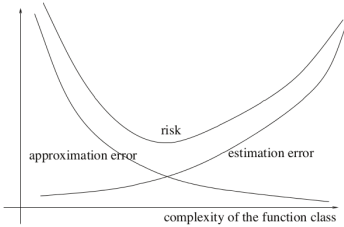


Figura obtida de Luxburg and Scholkopf, Statistical Learning Theory: Models, Concepts, and Results

Underfitting e Overfitting

• A partir dessa relação podemos definir:

- **Underfitting** – Se o subespaço F é pequeno, o erro de estimação é menor, mas o erro de aproximação é muito grande
- **Overfitting** – Se o subespaço F é grande, então o erro de estimação é muito grande, enquanto o erro de aproximação é pequeno
- O melhor para aprendizado de máquina é dado no ponto de equilíbrio
- Observe que a figura abaixo define o tamanho do subespaço F em termos da complexidade das funções contidas em F

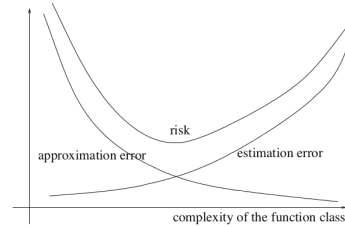


Figura obtida de Luxburg and Scholkopf, Statistical Learning Theory: Models, Concepts, and Results

Underfitting e Overfitting

• A partir dessa relação podemos definir:

- **Underfitting** – Se o subespaço F é pequeno, o erro de estimação é menor, mas o erro de aproximação é muito grande
- **Overfitting** – Se o subespaço F é grande, então o erro de estimação é muito grande, enquanto o erro de aproximação é pequeno
- O melhor para aprendizado de máquina é dado no ponto de equilíbrio
- Observe que a figura abaixo define o tamanho do subespaço F em termos da complexidade das funções contidas em F

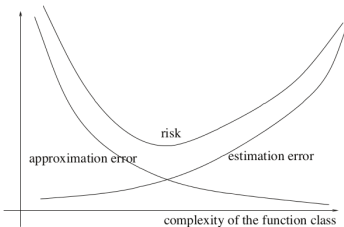
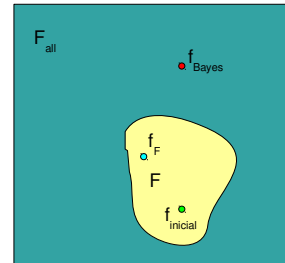


Figura obtida de Luxburg and Scholkopf, Statistical Learning Theory: Models, Concepts, and Results

Sobre viéses

- Algumas observações sobre viéses em aprendizado de máquina supervisionado



Sobre viéses

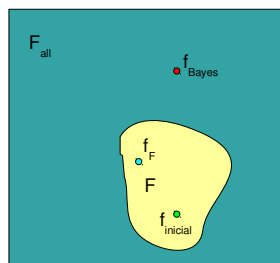
- Algumas observações sobre viéses em aprendizado de máquina supervisionado

Os viéses de um algoritmo de aprendizado são influenciados por:

1) **Viés de hipóteses** – Defina a linguagem de representação das hipóteses ou modelos. Exemplos: if-then rules, árvores de decisão, redes neurais artificiais, etc.

2) **Viés de preferência** – condições pelas quais o algoritmo prefere um classificador em relação a outro

3) **Viés de busca** – heurística ou critério de busca no subespaço F



Sobre viéses

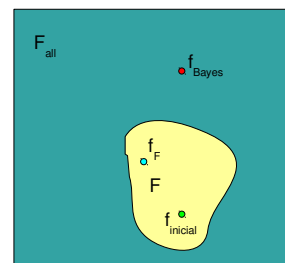
- Algumas observações sobre viéses em aprendizado de máquina supervisionado

Os viéses de um algoritmo de aprendizado são influenciados por:

1) **Viés de hipóteses** – Defina a linguagem de representação das hipóteses ou modelos. Exemplos: if-then rules, árvores de decisão, redes neurais artificiais, etc.

2) **Viés de preferência** – condições pelas quais o algoritmo prefere um classificador em relação a outro

3) **Viés de busca** – heurística ou critério de busca no subespaço F



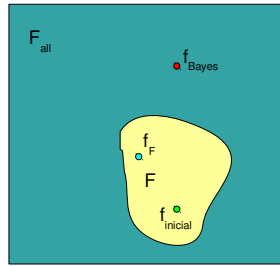
- Algumas observações sobre vieses em aprendizado de máquina supervisionado

Os vieses de um algoritmo de aprendizado são influenciados por:

1) **Viés de hipóteses** – Define a linguagem de representação das hipóteses ou modelos. Exemplos: if-then rules, árvores de decisão, redes neurais artificiais, etc.

2) **Viés de preferência** – condições pelas quais o algoritmo prefere um classificador em relação a outro

3) **Viés de busca** – heurística ou critério de busca no subespaço F



$$R(f) < R(g)$$

Começando a formalização da TAE

- Podemos chegar ao melhor classificador utilizando o Risco Esperado:

$$R(f) < R(g)$$

- Observe que as definições de **Consistência** anteriores utilizam o Risco Esperado para um classificador
 - Mas não temos como calcular esse Risco
- No entanto, o Risco Empírico é nosso principal estimador do Risco Esperado de um certo classificador
 - Mas como confiar no Risco Empírico se pode haver overfitting do modelo? Ex: Classificador com memória.
 - Daqui para frente iremos utilizar o Risco Empírico como forma de verificar consistência universal para um algoritmo com respeito a um subespaço F de classificadores

Começando a formalização da TAE

- Podemos chegar ao melhor classificador utilizando o Risco Esperado:

$$R(f) < R(g)$$

- Observe que as definições de **Consistência** anteriores utilizam o Risco Esperado para um classificador
 - Mas não temos como calcular esse Risco
- No entanto, o Risco Empírico é nosso principal estimador do Risco Esperado de um certo classificador
 - Mas como confiar no Risco Empírico se pode haver overfitting do modelo? Ex: Classificador com memória.
 - Daqui para frente iremos utilizar o Risco Empírico como forma de verificar consistência universal para um algoritmo com respeito a um subespaço F de classificadores

Começando a formalização da TAE

- Podemos chegar ao melhor classificador utilizando o Risco Esperado:

$$R(f) < R(g)$$

- Observe que as definições de **Consistência** anteriores utilizam o Risco Esperado para um classificador
 - Mas não temos como calcular esse Risco
- No entanto, o Risco Empírico é nosso principal estimador do Risco Esperado de um certo classificador
 - Mas como confiar no Risco Empírico se pode haver overfitting do modelo? Ex: Classificador com memória.
 - Daqui para frente iremos utilizar o Risco Empírico como forma de verificar consistência universal para um algoritmo com respeito a um subespaço F de classificadores

Começando a formalização da TAE

- Podemos chegar ao melhor classificador utilizando o Risco Esperado:

$$R(f) < R(g)$$

- Observe que as definições de **Consistência** anteriores utilizam o Risco Esperado para um classificador
 - Mas não temos como calcular esse Risco
- No entanto, o Risco Empírico é nosso principal estimador do Risco Esperado de um certo classificador
 - Mas como confiar no Risco Empírico se pode haver overfitting do modelo? Ex: Classificador com memória.
 - Daqui para frente iremos utilizar o Risco Empírico como forma de verificar consistência universal para um algoritmo com respeito a um subespaço F de classificadores

Começando a formalização da TAE

- Podemos chegar ao melhor classificador utilizando o Risco Esperado:

$$R(f) < R(g)$$

- Observe que as definições de **Consistência** anteriores utilizam o Risco Esperado para um classificador
 - Mas não temos como calcular esse Risco
- No entanto, o Risco Empírico é nosso principal estimador do Risco Esperado de um certo classificador
 - Mas como confiar no Risco Empírico se pode haver overfitting do modelo? Ex: Classificador com memória.
 - Daqui para frente iremos utilizar o Risco Empírico como forma de verificar consistência universal para um algoritmo com respeito a um subespaço F de classificadores

Começando a formalização da TAE

- Podemos chegar ao melhor classificador utilizando o Risco Esperado:

$$R(f) < R(g)$$

- Observe que as definições de **Consistência** anteriores utilizam o Risco Esperado para um classificador
 - Mas não temos como calcular esse Risco
- No entanto, o Risco Empírico é nosso principal estimador do Risco Esperado de um certo classificador
 - Mas como confiar no Risco Empírico se pode haver overfitting do modelo? Ex: Classificador com memória.
 - Daqui para frente iremos utilizar o Risco Empírico como forma de verificar consistência universal para um algoritmo com respeito a um subespaço F de classificadores

Assim surge o Princípio de Minimização do Risco Empírico

Princípio da Minimização do Risco Empírico

$$R(f) = E(l(x, y, f(x)))$$

Princípio da Minimização do Risco Empírico

- O problema de aprendizado de máquina consiste em encontrar um dado classificador f que minimize o Risco Esperado:

$$R(f) = E(l(x, y, f(x)))$$

- No entanto:
 - Não temos como computar R(f), pois não temos acesso à distribuição conjunta de probabilidades P
- Logo, poderíamos aproximar o Risco Esperado por meio do Risco Empírico e minimizá-lo:

Princípio da Minimização do Risco Empírico

- O problema de aprendizado de máquina consiste em encontrar um dado classificador f que minimize o Risco Esperado:

$$R(f) = E(l(x, y, f(x)))$$

- No entanto:
 - Não temos como computar R(f), pois não temos acesso à distribuição conjunta de probabilidades P
- Logo, poderíamos aproximar o Risco Esperado por meio do Risco Empírico e minimizá-lo:

$$R_{emp}(f) = \frac{1}{n} \sum_{i=1}^n l(x_i, y_i, f(x_i))$$

Princípio da Minimização do Risco Empírico

Princípio da Minimização do Risco Empírico

- Dessa maneira, assumindo:
 - Um subespaço de funções F
 - Uma função de perda
- Vapnik propõe a aproximação do Risco Esperado por meio do Risco Empírico e, assim, busca por um classificador f_n tal que:

$$f_n = \operatorname{argmin}_{f \in F} R_{emp}(f)$$

- A motivação de Vapnik para isso foi a Lei dos Grandes Números...

Princípio da Minimização do Risco Empírico

- Dessa maneira, assumindo:
 - Um subespaço de funções F
 - Uma função de perda
- Vapnik propõe a aproximação do Risco Esperado por meio do Risco Empírico e, assim, busca por um classificador f_n tal que:

$$f_n = \operatorname{argmin}_{f \in F} R_{emp}(f)$$

- A motivação de Vapnik para isso foi a Lei dos Grandes Números...

Princípio da Minimização do Risco Empírico

Princípio da Minimização do Risco Empírico

- Vapnik percebeu a possibilidade de considerar a **Lei dos Grandes Números**
 - Um importante Teorema da Estatística
- Segundo essa lei, assumindo que os dados foram amostrados de maneira independente e identicamente distribuídos (**iid**) a partir de alguma distribuição de probabilidades P , a média de uma amostra converge para a média populacional de P conforme o tamanho da amostra aumenta:

Princípio da Minimização do Risco Empírico

Princípio da Minimização do Risco Empírico

- Vapnik percebeu a possibilidade de considerar a **Lei dos Grandes Números**
 - Um importante Teorema da Estatística
- Segundo essa lei, assumindo que os dados foram amostrados de maneira independente e identicamente distribuídos (**iid**) a partir de alguma distribuição de probabilidades P , a média de uma amostra converge para a média populacional de P conforme o tamanho da amostra aumenta:

$$\frac{1}{n} \sum_{i=1}^n \xi_i \rightarrow E(\xi) \quad \text{para } n \rightarrow \infty$$

- Da mesma maneira Vapnik assumiu que o Risco Empírico de um classificador f converge para o Risco Esperado de f conforme o tamanho da amostra tende ao infinito:

$$R_{emp}(f) = \frac{1}{n} \sum_{i=1}^n l(x_i, y_i, f(x_i)) \rightarrow E(l(x, y, f(x)))$$

Para $n \rightarrow \infty$

Princípio da Minimização do Risco Empírico

$$P\left(\left|\frac{1}{n} \sum_{i=1}^n \xi_i - E(\xi)\right| \geq \epsilon\right) \leq 2 \exp(-2n\epsilon^2)$$

Sendo ξ_i valores aleatórios no intervalo $[0, 1]$

Princípio da Minimização do Risco Empírico

- Chernoff (1952) propôs uma desigualdade, estendida por Hoeffding (1963), que caracteriza quão bem uma média empírica se aproxima do valor esperado na forma:

$$P\left(\left|\frac{1}{n} \sum_{i=1}^n \xi_i - E(\xi)\right| \geq \epsilon\right) \leq 2 \exp(-2n\epsilon^2)$$

Sendo ξ_i valores aleatórios no intervalo $[0, 1]$

- Segundo este Teorema:
 - A probabilidade da média amostral desviar mais que um epsilon do valor esperado é limitada por uma pequena quantidade definida ao lado direito da desigualdade
 - Note que conforme o valor de n aumenta, torna-se menor o valor do lado direito

Princípio da Minimização do Risco Empírico

- Assim, pode-se aplicar a desigualdade de Hoeffding:

$$P\left(\left|\frac{1}{n} \sum_{i=1}^n \xi_i - E(\xi)\right| \geq \epsilon\right) \leq 2 \exp(-2n\epsilon^2)$$

Sendo ξ_i valores aleatórios no intervalo $[0, 1]$

- Para obter um limite que define o quanto, para um certo classificador f, o Risco Empírico se aproxima do Risco Esperado:

$$P(|R_{emp}(f) - R(f)| \geq \epsilon) \leq 2 \exp(-2n\epsilon^2)$$

Princípio da Minimização do Risco Empírico

- Assim, pode-se aplicar a desigualdade de Hoeffding:

No entanto, esse Teorema assume:

- Um valor para n suficientemente grande
- OK

- Para
- class
- Espera

$$P(|R_{emp}(f) - R(f)| \geq \epsilon) \leq 2 \exp(-2n\epsilon^2)$$

Princípio da Minimização do Risco Empírico

- Assim, pode-se aplicar a desigualdade de Hoeffding:

No entanto, esse Teorema assume:

- Essa função não pode ser obtida com base no conjunto de treinamento, i.e., ela deve ser independente do conjunto de treinamento
- Problema

- Pa
- Q

Princípio da Minimização do Risco Empírico

- Assim, pode-se aplicar a desigualdade de Hoeffding:

No entanto, esse Teorema assume:

No entanto, esse Teorema assume:

- Essa função deve ser fixa!
- Problema

- Pa
- Q

Princípio da Minimização do Risco Empírico

Princípio da Minimização do Risco Empírico

- No entanto, há restrições relevantes sobre o limite de Chernoff:
 - **Ele é válido somente para uma função ou classificador f fixo definido**
 - **O classificador deve ser escolhido de forma independente do conjunto de treinamento**
- Contudo, em aprendizado de máquina:
 - A função f é obtida com base no conjunto de treinamento
 - Saímos de uma função inicial e desejamos convergir para a melhor dentro do subespaço definido pelo Algoritmo de Classificação
 - Isso, a priori, invalida o uso da Lei dos Grandes Números para provar que o Risco Empírico pode ser um bom estimador para o Risco Esperado

Princípio da Minimização do Risco Empírico

Princípio da Minimização do Risco Empírico

- No entanto, há restrições relevantes sobre o limite de Chernoff:
 - **Ele é válido somente para uma função ou classificador f fixo definido**
 - **O classificador deve ser escolhido de forma independente do conjunto de treinamento**
 - Contudo, em aprendizado de máquina:
 - A função f é obtida com base no conjunto de treinamento
 - Saímos de uma função inicial e desejamos convergir para a melhor dentro do subespaço definido pelo Algoritmo de Classificação
 - Isso, a priori, invalida o uso da Lei dos Grandes Números para provar que o Risco Empírico pode ser um bom estimador para o Risco Esperado
- No entanto, há restrições relevantes sobre o limite de Chernoff:
 - **Ele é válido somente para uma função ou classificador f fixo definido**
 - **O classificador deve ser escolhido de forma independente do conjunto de treinamento**
 - Contudo, em aprendizado de máquina:
 - A função f é obtida com base no conjunto de treinamento
 - Saímos de uma função inicial e desejamos convergir para a melhor dentro do subespaço definido pelo Algoritmo de Classificação
 - Isso, a priori, invalida o uso da Lei dos Grandes Números para provar que o Risco Empírico pode ser um bom estimador para o Risco Esperado

Princípio da Minimização do Risco Empírico

Princípio da Minimização do Risco Empírico

- No entanto, há restrições relevantes sobre o limite de Chernoff:
 - **Ele é válido somente para uma função ou classificador f fixo definido**
 - **O classificador deve ser escolhido de forma independente do conjunto de treinamento**
 - Contudo, em aprendizado de máquina:
 - A função f é obtida com base no conjunto de treinamento
 - Saímos de uma função inicial e desejamos convergir para a melhor dentro do subespaço definido pelo Algoritmo de Classificação
 - Isso, a priori, invalida o uso da Lei dos Grandes Números para provar que o Risco Empírico pode ser um bom estimador para o Risco Esperado
- No entanto, há restrições relevantes sobre o limite de Chernoff:
 - **Ele é válido somente para uma função ou classificador f fixo definido**
 - **O classificador deve ser escolhido de forma independente do conjunto de treinamento**
 - Contudo, em aprendizado de máquina:
 - A função f é obtida com base no conjunto de treinamento
 - Saímos de uma função inicial e desejamos convergir para a melhor dentro do subespaço definido pelo Algoritmo de Classificação
 - Isso, a priori, invalida o uso da Lei dos Grandes Números para provar que o Risco Empírico pode ser um bom estimador para o Risco Esperado

Princípio da Minimização do Risco Empírico

- No entanto, há restrições relevantes sobre o limite de Chernoff:
 - Ele é válido somente para uma função ou classificador f fixo definido**
 - O princípio da minimização do risco empírico é inconsistente**
- O que levaria à inconsistência do Princípio da Minimização do Risco Empírico**
- Mas Vapnik não desistiu...**
- Isso, a priori, invalida o uso da Lei dos Grandes Números para provar que o Risco Empírico pode ser um bom estimador para o Risco Esperado

Inconsistência do Princípio da Minimização do Risco Empírico

- Por exemplo, considere o espaço de exemplos versus rótulos formado por uma função determinística:

$$Y = \begin{cases} -1 & \text{if } X < 0.5 \\ 1 & \text{if } X \geq 0.5 \end{cases}$$

- Assuma que produzimos um classificador f que produz erro zero no conjunto de treinamento, pois simplesmente memoriza as respostas Y para as entradas X na forma:

$$f_n(X) = \begin{cases} Y_i & \text{if } X = X_i \text{ for some } i = 1, \dots, n \\ 1 & \text{otherwise.} \end{cases}$$

Inconsistência do Princípio da Minimização do Risco Empírico

Inconsistência do Princípio da Minimização do Risco Empírico

- O Risco Empírico desse classificador f no conjunto de treinamento é zero
 - No entanto, para exemplos nunca vistos, ele sempre atribui rótulo ou classe igual a 1
 - Considerando que metade dos exemplos terá essa classe, então ele acerta em 50% dos casos e erra em outros 50% dos casos
 - Sendo assim, o Risco Empírico de f não se aproxima do Risco Esperado para f nesse cenário
 - Ou seja, o princípio da minimização do Risco Empírico é inconsistente para este caso
 - Assumindo que o classificador foi obtido com base no conjunto de treinamento**

Inconsistência do Princípio da Minimização do Risco Empírico

Inconsistência do Princípio da Minimização do Risco Empírico

- O Risco Empírico desse classificador f no conjunto de treinamento é zero
 - No entanto, para exemplos nunca vistos, ele sempre atribui rótulo ou classe igual a 1
 - Considerando que metade dos exemplos terá essa classe, então ele acerta em 50% dos casos e erra em outros 50% dos casos
 - Sendo assim, o Risco Empírico de f não se aproxima do Risco Esperado para f nesse cenário
 - Ou seja, o princípio da minimização do Risco Empírico é inconsistente para este caso
 - Assumindo que o classificador foi obtido com base no conjunto de treinamento**
- O Risco Empírico desse classificador f no conjunto de treinamento é zero
 - No entanto, para exemplos nunca vistos, ele sempre atribui rótulo ou classe igual a 1
 - Considerando que metade dos exemplos terá essa classe, então ele acerta em 50% dos casos e erra em outros 50% dos casos
 - Sendo assim, o Risco Empírico de f não se aproxima do Risco Esperado para f nesse cenário
 - Ou seja, o princípio da minimização do Risco Empírico é inconsistente para este caso
 - Assumindo que o classificador foi obtido com base no conjunto de treinamento**

Inconsistência do Princípio da Minimização do Risco Empírico

- O Risco Empírico desse classificador f no conjunto de treinamento é zero
 - No entanto, para exemplos nunca vistos, ele sempre atribui rótulo ou classe igual a 1
 - Considerando que metade dos exemplos terá essa classe, então ele acerta em 50% dos casos e erra em outros 50% dos casos
 - Sendo assim, o Risco Empírico de f não se aproxima do Risco Esperado para f nesse cenário
 - Ou seja, o princípio da minimização do Risco Empírico é inconsistente para este caso
 - **Assumindo que o classificador foi obtido com base no conjunto de treinamento**

Inconsistência do Princípio da Minimização do Risco Empírico

Inconsistência do Princípio da Minimização do Risco Empírico

- Este é um exemplo de extremo **overfitting**
 - Ou seja, o classificador f é ótimo para o conjunto de treinamento, mas falha ao máximo para exemplos nunca vistos
 - Observe se temos 2 opções, 50% é obtido simplesmente “chutando” a classe -1 ou +1 para cada exemplo não visto
- No entanto, temos apenas o Risco Empírico para trabalhar e estimar o Risco Esperado:
 - **Poderíamos, então, de alguma forma resgatar o Princípio de Minimização do Risco Empírico e torná-lo um bom estimador para o Risco Esperado?**

Inconsistência do Princípio da Minimização do Risco Empírico

Inconsistência do Princípio da Minimização do Risco Empírico

- Poderíamos, então, de alguma forma resgatar o princípio de minimização do Risco Empírico e torná-lo um bom estimador para o Risco Esperado?
 - **Sim, mas para isso devemos avaliar a classe de funções ou subespaço F contido em F_{all} e que será considerado por nosso algoritmo de aprendizado**
 - **O Princípio deve ser consistente para toda função nesse espaço, pois todas podem ser escolhidas como classificador**
 - Se considerarmos um subespaço F que contenha a função de memorização dos exemplos de treinamento, então este princípio de minimização **NUNCA** irá funcionar!
 - Portanto, precisamos fazer restrições quanto ao subespaço F utilizado para estimar nosso classificador f
 - Caso contrário não podemos esperar por aprendizado
 - Felizmente temos maneiras de restringir o subespaço F

Inconsistência do Princípio da Minimização do Risco Empírico

- Poderíamos, então, de alguma forma resgatar o princípio de minimização do Risco Empírico e torná-lo um bom estimador para o Risco Esperado?
 - **Sim, mas para isso devemos avaliar a classe de funções ou subespaço F contido em F_{all} e que será considerado por nosso algoritmo de aprendizado**
 - **O Princípio deve ser consistente para toda função nesse espaço, pois todas podem ser escolhidas como classificador**
 - Se considerarmos um subespaço F que contenha a função de memorização dos exemplos de treinamento, então este princípio de minimização **NUNCA** irá funcionar!
 - Portanto, precisamos fazer restrições quanto ao subespaço F utilizado para estimar nosso classificador f
 - Caso contrário não podemos esperar por aprendizado
 - Felizmente temos maneiras de restringir o subespaço F

Inconsistência do Princípio da Minimização do Risco Empírico

- Poderíamos, então, de alguma forma resgatar o princípio de minimização do Risco Empírico e torná-lo um bom estimador para o Risco Esperado?
 - Sim, mas para isso devemos avaliar a classe de funções ou subespaço F contido em F_{all} e que será considerado por nosso algoritmo de aprendizado**
 - O Princípio deve ser consistente para toda função nesse espaço, pois todas podem ser escolhidas como classificador**
 - Se considerarmos um subespaço F que contenha a função de memorização dos exemplos de treinamento, então este princípio de minimização **NUNCA** irá funcionar!
 - Portanto, precisamos fazer restrições quanto ao subespaço F utilizado para estimar nosso classificador f
 - Caso contrário não podemos esperar por aprendizado
 - Felizmente temos maneiras de restringir o subespaço F

Inconsistência do Princípio da Minimização do Risco Empírico

- Poderíamos, então, de alguma forma resgatar o princípio de minimização do Risco Empírico e torná-lo um bom estimador para o Risco Esperado?
 - Sim, mas para isso devemos avaliar a classe de funções ou subespaço F contido em F_{all} e que será considerado por nosso algoritmo de aprendizado**
 - O Princípio deve ser consistente para toda função nesse espaço, pois todas podem ser escolhidas como classificador**
 - Se considerarmos um subespaço F que contenha a função de memorização dos exemplos de treinamento, então este princípio de minimização **NUNCA** irá funcionar!
 - Portanto, precisamos fazer restrições quanto ao subespaço F utilizado para estimar nosso classificador f
 - Caso contrário não podemos esperar por aprendizado
 - Felizmente temos maneiras de restringir o subespaço F

Tornando o Princípio da Minimização do Risco Empírico Consistente

Tornando o Princípio da Minimização do Risco Empírico Consistente

- As condições necessárias para tornar o Princípio da Minimização do Risco Empírico Consistente envolvem a **restrição no espaço de funções admissíveis**
 - Ou seja, a restrição do subespaço F que contém as funções ou possíveis classificadores que iremos considerar
- Para isso podemos considerar o **pior caso para todos os possíveis classificadores f pertencentes ao subespaço F**
 - Pois qualquer um desses classificadores pode ser escolhido por nosso algoritmo de aprendizado
 - Para isso podemos verificar se para toda função f contida no subespaço F , o Risco Empírico converge para Risco Esperado conforme o número de exemplos tende ao infinito**

Tornando o Princípio da Minimização do Risco Empírico Consistente

- As condições necessárias para tornar o Princípio da Minimização do Risco Empírico Consistente envolvem a **restrição no espaço de funções admissíveis**
 - Ou seja, a restrição do subespaço F que contém as funções ou possíveis classificadores que iremos considerar
- Para isso podemos considerar o **pior caso para todos os possíveis classificadores f pertencentes ao subespaço F**
 - Pois qualquer um desses classificadores pode ser escolhido por nosso algoritmo de aprendizado
 - Para isso podemos verificar se para toda função f contida no subespaço F , o Risco Empírico converge para Risco Esperado conforme o número de exemplos tende ao infinito**

Se convergir para o pior caso, ou seja, a pior função, então ele converge para as demais no subespaço F

Tornando o Princípio da Minimização do Risco Empírico Consistente

- Uma maneira de garantirmos essa convergência para toda função f contida em F é por meio da **convergência uniforme sobre F** :
 - Assim, conforme n (número de exemplos de treinamento) aumenta, a diferença abaixo deve reduzir
- Consequentemente, para um certo valor de n , a desigualdade será válida para um epsilon pequeno:

$$|R_{emp}(f) - R(f)|$$

- Considerando toda f pertencente ao subespaço F

$$|R_{emp}(f) - R(f)| \leq \epsilon \text{ para toda função } f \in F, F \subset F_{all}$$

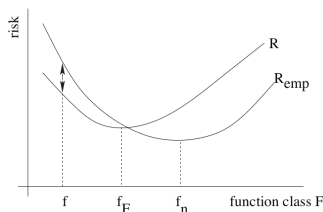


Figura obtida de Luxburg and Scholkopf, Statistical Learning Theory: Models, Concepts, and Results

- Isso significa que a distância entre as curvas do Risco Empírico e do Risco Esperado para todas as funções f no subespaço F nunca é maior que um dado epsilon

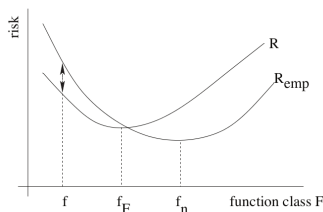


Figura obtida de Luxburg and Scholkopf, Statistical Learning Theory: Models, Concepts, and Results

- Matematicamente, podemos representar esse conceito pelo supremo:

$$\sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)| \leq \varepsilon.$$

- Isso significa que a distância entre as curvas do Risco Empírico e do Risco Esperado para todas as funções f no subespaço F nunca é maior que um dado epsilon

Observe que o supremo considera o pior caso!

- Matematicamente, podemos representar esse conceito pelo supremo:

$$\sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)| \leq \varepsilon.$$

$$|R(f) - R_{\text{emp}}(f)| \leq \sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)|$$

- Assim, se a desigualdade abaixo é verdadeira, então toda diferença entre Risco Empírico e Risco Esperado de toda função f em F está limitada por esse supremo:

$$|R(f) - R_{\text{emp}}(f)| \leq \sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)|$$

- A partir disso podemos concluir que:

$$P(|R(f_n) - R_{\text{emp}}(f_n)| \geq \varepsilon) \leq P(\sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)| \geq \varepsilon)$$

$$P(|R(f_n) - R_{\text{emp}}(f_n)| \geq \varepsilon) \leq P(\sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)| \geq \varepsilon)$$

Tornando o Princípio da Minimização do Risco Empírico Consistente

- À direita do termo a seguir temos exatamente a situação em que a **Lei Uniforme dos Grandes Números trata, i.e., considerando um conjunto de funções fixas** (classificadores fixos):

$$P(|R(f_n) - R_{\text{emp}}(f_n)| \geq \varepsilon) \leq P(\sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)| \geq \varepsilon)$$

- Portanto, podemos dizer que a Lei dos Grandes Números é uniformemente válida sobre uma classe de funções (ou subespaço F) para todo epsilon maior que zero (lembre-se, epsilon é uma distância) na forma:

$$P(\sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)| \geq \varepsilon) \rightarrow 0 \text{ as } n \rightarrow \infty$$

- Sendo essa probabilidade um limitante superior para toda função f em F

Tornando o Princípio da Minimização do Risco Empírico Consistente

$$P(|R(f_n) - R_{\text{emp}}(f_n)| \geq \varepsilon) \leq P(\sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)| \geq \varepsilon)$$

Tornando o Princípio da Minimização do Risco Empírico Consistente

- Agora podemos usar a desigualdade abaixo:

$$P(|R(f_n) - R_{\text{emp}}(f_n)| \geq \varepsilon) \leq P(\sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)| \geq \varepsilon)$$

- **Para mostrar que a Lei Uniforme dos Grandes Números é válida para uma classe de funções F, ou seja, para um subespaço F com um conjunto restrito de funções**

- Logo, o princípio da minimização do Risco Empírico é consistente com respeito ao subespaço F
- Para isso considere a distância entre o Risco Esperado de um classificador qualquer contido no subespaço F versus o melhor deles em F

$$|R(f_n) - R(f_{\mathcal{F}})|$$

$$|R(f_n) - R(f_F)| \geq 0$$

Tornando o Princípio da Minimização do Risco Empírico Consistente

- É óbvio que:

$$|R(f_n) - R(f_F)| \geq 0$$

- Pois o Risco Esperado do melhor classificador no subespaço F é menor ou igual a qualquer outro classificador contido em F, logo:

$$|R(f_n) - R(f_F)| = R(f_n) - R(f_F)$$

- Vapnik adicionou termos relativos aos Riscos Empíricos na forma:

Tornando o Princípio da Minimização do Risco Empírico Consistente

- É óbvio que:

$$|R(f_n) - R(f_F)| \geq 0$$

- Pois o Risco Esperado do melhor classificador no subespaço F é menor ou igual a qualquer outro classificador contido em F, logo:

$$|R(f_n) - R(f_F)| = R(f_n) - R(f_F)$$

- Vapnik adicionou termos relativos aos Riscos Empíricos na forma:

$$R(f_n) - R(f_F) = R(f_n) - R_{\text{emp}}(f_n) + R_{\text{emp}}(f_n) - R_{\text{emp}}(f_F) + R_{\text{emp}}(f_F) - R(f_F)$$

Tornando o Princípio da Minimização do Risco Empírico Consistente

- Logo, Vapnik conclui que:

$$R(f_n) - R(f_F) = R(f_n) - R_{emp}(f_n) + R_{emp}(f_n) - R_{emp}(f_F) + R_{emp}(f_F) - R(f_F)$$

- é limitado pelo termo à direita:

$$R(f_n) - R_{emp}(f_n) + R_{emp}(f_n) - R_{emp}(f_F) + R_{emp}(f_F) - R(f_F) \leq R(f_n) - R_{emp}(f_n) + R_{emp}(f_F) - R(f_F)$$

Tornando o Princípio da Minimização do Risco Empírico Consistente

- Logo, Vapnik conclui que:

$$R(f_n) - R(f_F) = R(f_n) - R_{emp}(f_n) + R_{emp}(f_n) - R_{emp}(f_F) + R_{emp}(f_F) - R(f_F)$$

- é limitado pelo termo à direita:

$$R(f_n) - R_{emp}(f_n) + R_{emp}(f_n) - R_{emp}(f_F) + R_{emp}(f_F) - R(f_F) \leq R(f_n) - R_{emp}(f_n) + R_{emp}(f_F) - R(f_F)$$

- Pois, por definição, o Risco Empírico do melhor classificador em F é menor que de qualquer outro em F, exceto ele mesmo, conforme n aumenta, i.e.:

$$R_{emp}(f_n) - R_{emp}(f_F) \leq 0$$

Tornando o Princípio da Minimização do Risco Empírico Consistente

$$R(f_n) - R_{emp}(f_n) + R_{emp}(f_n) - R_{emp}(f_F) + R_{emp}(f_F) - R(f_F) \leq R(f_n) - R_{emp}(f_n) + R_{emp}(f_F) - R(f_F)$$

Tornando o Princípio da Minimização do Risco Empírico Consistente

- Vapnik ainda trabalhou sobre a desigualdade:

$$R(f_n) - R_{emp}(f_n) + R_{emp}(f_n) - R_{emp}(f_F) + R_{emp}(f_F) - R(f_F) \leq R(f_n) - R_{emp}(f_n) + R_{emp}(f_F) - R(f_F)$$

- Observe que ela tem, do lado direito, a distância entre o Risco Empírico e Esperado para a função f_n e para a função f_F
- Vapnik considerou a pior função f_n contida em F**, ou seja, aquela com maior distância entre Risco Empírico e Esperado para, assim, obter outro limitante:

Tornando o Princípio da Minimização do Risco Empírico Consistente

- Vapnik ainda trabalhou sobre a desigualdade:

$$R(f_n) - R_{emp}(f_n) + R_{emp}(f_n) - R_{emp}(f_F) + R_{emp}(f_F) - R(f_F) \leq R(f_n) - R_{emp}(f_n) + R_{emp}(f_F) - R(f_F)$$

- Observe que ela tem, do lado direito, a distância entre o Risco Empírico e Esperado para a função f_n e para a função f_F
- Vapnik considerou a pior função f_n contida em F**, ou seja, aquela com maior distância entre Risco Empírico e Esperado para, assim, obter outro limitante:

$$R(f_n) - R_{emp}(f_n) + R_{emp}(f_F) - R(f_F) \leq 2 \sup_{f \in F} |R(f) - R_{emp}(f)|$$

Tornando o Princípio da Minimização do Risco Empírico Consistente

- Vapnik ainda trabalhou sobre a desigualdade:

$$R(f_n) - R_{emp}(f_n) + R_{emp}(f_n) - R_{emp}(f_F) + R_{emp}(f_F) - R(f_F) \leq R(f_n) - R_{emp}(f_n) + R_{emp}(f_F) - R(f_F)$$

- Observe que o termo: $2 \sup_{f \in F} |R(f) - R_{emp}(f)|$ surge como limitante superior para a pior Função do subespaço F
- Vapnik considerou a pior função f_n contida em F**, ou seja, aquela com maior distância entre Risco Empírico e Esperado para, assim, obter outro limitante:

$$R(f_n) - R_{emp}(f_n) + R_{emp}(f_F) - R(f_F) \leq 2 \sup_{f \in F} |R(f) - R_{emp}(f)|$$

Tornando o Princípio da Minimização do Risco Empírico Consistente

- Vapnik ainda trabalhou sobre a desigualdade:

$R(f_n) - R(f_{\mathcal{F}}) + R(f_{\mathcal{F}}) - R(f_F) \leq 2 \sup_{f \in F} |R(f) - R_{emp}(f)|$

Observe que o termo:

Observe que Vapnik constantemente busca por limitantes superiores em seu trabalho

Por que???

Tornando o Princípio da Minimização do Risco Empírico Consistente

- Vapnik ainda trabalhou sobre a desigualdade:

$R(f_n) - R(f_{\mathcal{F}}) + R(f_{\mathcal{F}}) - R(f_F) \leq 2 \sup_{f \in F} |R(f) - R_{emp}(f)|$

Observe que o termo:

Se garantimos algo para o limitante superior, garantimos também para o lado esquerdo, i.e., aquele que é limitado por um termo à direita

Tornando o Princípio da Minimização do Risco Empírico Consistente

$$|R(f_n) - R(f_{\mathcal{F}})|$$

Tornando o Princípio da Minimização do Risco Empírico Consistente

- Vapnik volta, então, no erro de estimação:

$$|R(f_n) - R(f_{\mathcal{F}})|$$

- E considera:

$$R(f_n) - R_{emp}(f_n) + R_{emp}(f_F) - R(f_F) \leq 2 \sup_{f \in F} |R(f) - R_{emp}(f)|$$

- Para obter:

Tornando o Princípio da Minimização do Risco Empírico Consistente

- Vapnik volta, então, no erro de estimação:

$$|R(f_n) - R(f_{\mathcal{F}})|$$

- E considera:

$$R(f_n) - R_{emp}(f_n) + R_{emp}(f_F) - R(f_F) \leq 2 \sup_{f \in F} |R(f) - R_{emp}(f)|$$

- Para obter:

$$P(|R(f_n) - R(f_{\mathcal{F}})| \geq \varepsilon) \leq P(\sup_{f \in F} |R(f) - R_{emp}(f)| \geq \varepsilon/2)$$

Tornando o Princípio da Minimização do Risco Empírico Consistente

- Vapnik volta, então, no erro de estimação:

$$|R(f_n) - R(f_{\mathcal{F}})|$$

- E consi

$$R(f_n) - R_{emp}(f_n) + R_{emp}(f_F) - R(f_F) \leq 2 \sup_{f \in F} |R(f) - R_{emp}(f)|$$

- Para

$$P(|R(f_n) - R(f_{\mathcal{F}})| \geq \varepsilon) \leq P(\sup_{f \in F} |R(f) - R_{emp}(f)| \geq \varepsilon/2)$$

Assim o termo do lado direito, que considera apenas a pior função em F é suficiente para limitar a distância entre o pior e o melhor classificador contido no subespaço F

Tornando o Princípio da Minimização do Risco Empírico Consistente

$$P(|R(f_n) - R(f_{\mathcal{F}})| \geq \varepsilon) \leq P(\sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)| \geq \varepsilon/2)$$

Tornando o Princípio da Minimização do Risco Empírico Consistente

- Assim, se o supremo (para o pior classificador) da diferença dos Riscos Empírico e Esperado para toda função f em F convergir para zero conforme n aumenta:

$$P(|R(f_n) - R(f_{\mathcal{F}})| \geq \varepsilon) \leq P(\sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)| \geq \varepsilon/2)$$

- É condição suficiente para consistência do Princípio de Minimização do Risco Empírico
- Conclusões Importantes:**
 - Precisamos criar um viés, ou seja, escolher um espaço de funções para tornar o princípio válido
 - A condição de convergência uniforme depende crucialmente da definição de um subconjunto de funções
 - Sem viés não há garantias de aprendizado

Tornando o Princípio da Minimização do Risco Empírico Consistente

- Assim, se o supremo (para o pior classificador) da diferença dos Riscos Empírico e Esperado para toda função f em F convergir para zero conforme n aumenta:

$$P(|R(f_n) - R(f_{\mathcal{F}})| \geq \varepsilon) \leq P(\sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)| \geq \varepsilon/2)$$

$$\sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)|$$

- É condição suficiente para consistência do Princípio de Minimização do Risco Empírico
- Conclusões Importantes:**
 - Precisamos criar um viés, ou seja, escolher um espaço de funções para tornar o princípio válido
 - A condição de convergência uniforme depende crucialmente da definição de um subconjunto de funções
 - Sem viés não há garantias de aprendizado

Tornando o Princípio da Minimização do Risco Empírico Consistente

Conclusões Importantes:

- Quão maior o subespaço F , maior o valor de:

$$\sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)|$$

- Consequentemente, conforme F é maior, torna-se mais difícil satisfazer a Lei Uniforme dos Grandes Números
 - Em palavras simples:
 - Conforme o subespaço F aumenta, torna-se mais difícil atingir a consistência do PMRE
 - Subespaços F menores são mais fáceis de atingir consistência para o PMRE
 - Qual a importância da consistência do PMRE?**
 - Se consistente, o Risco Empírico se aproxima do Risco Esperado

Tornando o Princípio da Minimização do Risco Empírico Consistente

Conclusões Importantes:

- Quão maior o subespaço F , maior o valor de:

$$\sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)|$$

- Consequentemente, conforme F é maior, torna-se mais difícil satisfazer a Lei Uniforme dos Grandes Números
 - Em palavras simples:
 - Conforme o subespaço F aumenta, torna-se mais difícil atingir a consistência do PMRE
 - Subespaços F menores são mais fáceis de atingir consistência para o PMRE
 - Qual a importância da consistência do PMRE?**
 - Se consistente, o Risco Empírico se aproxima do Risco Esperado

- **Conclusões Importantes:**

- Quanto maior o subespaço F , maior o valor de:

$$\sup_{f \in F} |R(f) - R_{\text{emp}}(f)|$$

- Consequentemente, conforme F é maior, torna-se mais difícil satisfazer a Lei Uniforme dos Grandes Números
 - Em palavras simples:
 - Conforme o subespaço F aumenta, torna-se mais difícil atingir a consistência do PMRE
 - Subespaços F menores são mais fáceis de atingir consistência para o PMRE
 - **Qual a importância da consistência do PMRE?**
 - **Se consistente, o Risco Empírico se aproxima do Risco Esperado**

- **Conclusões Importantes:**

- Quanto maior o subespaço F , maior o valor de:

• **Além disso, sendo o PMRE consistente, garantimos aprendizado!**

Não quer dizer que não ocorra aprendizado caso o princípio seja inconsistente!!

- Subespaços F menores são mais fáceis de atingir consistência para o PMRE
 - **Qual a importância da consistência do PMRE?**
 - **Se consistente, o Risco Empírico se aproxima do Risco Esperado**

Propriedades do subespaço F para Garantirmos Convergência UniformePropriedades do subespaço F para Garantirmos Convergência Uniforme

- Agora que verificamos que há condições para que o Princípio de Minimização do Risco Empírico seja consistente:
 - Apesar de tudo que mostramos até agora ser interessante devido aos aspectos teóricos envolvidos:
 - Na prática não ajuda muito
 - **Para termos resultados práticos precisamos definir características/propriedades do subespaço de funções F para o qual esse Princípio garante convergência uniforme**
 - Ou seja, quais propriedades o subespaço F de funções deve respeitar para que o Risco Empírico seja um bom estimador do Risco Esperado conforme n aumenta?

Propriedades do subespaço F para Garantirmos Convergência UniformePropriedades do subespaço F para Garantirmos Convergência Uniforme

- Agora que verificamos que há condições para que o Princípio de Minimização do Risco Empírico seja consistente:
 - Apesar de tudo que mostramos até agora ser interessante devido aos aspectos teóricos envolvidos:
 - Na prática não ajuda muito
 - **Para termos resultados práticos precisamos definir características/propriedades do subespaço de funções F para o qual esse Princípio garante convergência uniforme**
 - Ou seja, quais propriedades o subespaço F de funções deve respeitar para que o Risco Empírico seja um bom estimador do Risco Esperado conforme n aumenta?

- Agora que verificamos que há condições para que o Princípio de Minimização do Risco Empírico seja consistente:
 - Apesar de tudo que mostramos até agora ser interessante devido aos aspectos teóricos envolvidos:
 - Na prática não ajuda muito
 - **Para termos resultados práticos precisamos definir características/propriedades do subespaço de funções F para o qual esse Princípio garante convergência uniforme**
 - Ou seja, quais propriedades o subespaço F de funções deve respeitar para que o Risco Empírico seja um bom estimador do Risco Esperado conforme n aumenta?

$$P(|\frac{1}{n} \sum_{i=1}^n \xi_i - E(\xi)| \geq \epsilon) \leq 2 \exp(-2n\epsilon^2)$$

Sendo ξ_i valores aleatórios no intervalo $[0, 1]$

- Antes voltemos para a Lei dos Grandes Números cujo limite superior é dado por Chernoff:

$$P(|\frac{1}{n} \sum_{i=1}^n \xi_i - E(\xi)| \geq \epsilon) \leq 2 \exp(-2n\epsilon^2)$$

Sendo ξ_i valores aleatórios no intervalo $[0, 1]$

- Considerando que o subespaço F é finito e contém as seguintes funções $F = \{f_1, \dots, f_m\}$. Cada uma dessas funções satisfaz a Lei dos Grandes Números na forma do limite de Chernoff:

$$P(|R(f_i) - R_{emp}(f_i)| \geq \epsilon) \leq 2 \exp(-2n\epsilon^2)$$

- O que vale para cada função individual contida em F

- Antes voltemos para a Lei dos Grandes Números cujo limite superior é dado por Chernoff:

$$P(|\frac{1}{n} \sum_{i=1}^n \xi_i - E(\xi)| \geq \epsilon) \leq 2 \exp(-2n\epsilon^2)$$

As funções nesse subespaço respeitam a Lei dos Grandes Números, considerando que nenhuma foi escolhida com base nos exemplos de treinamento

- Com as seguintes funções $F = \{f_1, \dots, f_m\}$ satisfaz a Lei dos Grandes Números na forma de Chernoff:

$$P(|R(f_i) - R_{emp}(f_i)| \geq \epsilon) \leq 2 \exp(-2n\epsilon^2)$$

- O que vale para cada função individual contida em F

- No entanto, precisamos de algo que valha para todas as funções contidas em F e não apenas para cada função individual
- Para isso Vapnik reescreveu:

$$P(|R(f_i) - R_{emp}(f_i)| \geq \epsilon) \leq 2 \exp(-2n\epsilon^2)$$

- Na forma:

- E em seguida encontrou um limite superior em termos de probabilidades:

- No entanto, precisamos de algo que valha para todas as funções contidas em F e não apenas para cada função individual
- Para isso Vapnik reescreveu:

$$P(|R(f_i) - R_{emp}(f_i)| \geq \epsilon) \leq 2 \exp(-2n\epsilon^2)$$

- Na forma:

$$P(\sup_{f \in F} |R(f) - R_{emp}(f)| \geq \epsilon) = P(|R(f_1) - R_{emp}(f_1)| \geq \epsilon \text{ or } \dots \text{ or } |R(f_m) - R_{emp}(f_m)| \geq \epsilon)$$

- E em seguida encontrou um limite superior em termos de probabilidades:

- No entanto, precisamos de algo que valha para todas as funções contidas em F e não apenas para cada função individual
- Para isso Vapnik reescreveu:

$$P(|R(f_i) - R_{emp}(f_i)| \geq \epsilon) \leq 2 \exp(-2n\epsilon^2)$$

- Na forma:

$$P(\sup_{f \in F} |R(f) - R_{emp}(f)| \geq \epsilon) = P(|R(f_1) - R_{emp}(f_1)| \geq \epsilon \text{ or } \dots \text{ or } |R(f_m) - R_{emp}(f_m)| \geq \epsilon)$$

- E em seguida encontrou um limite superior em termos de probabilidades:

$$P(|R(f_1) - R_{emp}(f_1)| \geq \epsilon \text{ or } \dots \text{ or } |R(f_m) - R_{emp}(f_m)| \geq \epsilon) \leq \sum_{i=1}^m P(|R(f_i) - R_{emp}(f_i)| \geq \epsilon)$$

$$\sum_{i=1}^m P(|R(f_i) - R_{emp}(f_i)| \geq \epsilon) \leq 2m \exp(-2n\epsilon^2)$$

- Finalmente obtemos um limite usando Chernoff para todo o conjunto de funções contidas no subespaço F na forma:

$$\sum_{i=1}^m P(|R(f_i) - R_{emp}(f_i)| \geq \epsilon) \leq 2m \exp(-2n\epsilon^2)$$

- Logo encontramos um limite para a convergência uniforme considerando todas as **m** funções contidas no subespaço **finito** F na forma:

$$P(\sup_{f \in F} |R(f) - R_{emp}(f)| \geq \epsilon) \leq 2m \exp(-2n\epsilon^2)$$

- Observe que se o subespaço F for finito, **m** torna-se uma constante e a convergência uniforme ainda ocorre conforme n aumenta
- Prova-se, assim, que o princípio de Minimização do Risco Empírico sobre um subespaço finito de funções F é consistente
 - Mas como fica se o subespaço F for infinito?

- Finalmente obtemos um limite usando Chernoff para todo o conjunto de funções contidas no subespaço F na forma:

$$\sum_{i=1}^m P(|R(f_i) - R_{emp}(f_i)| \geq \epsilon) \leq 2m \exp(-2n\epsilon^2)$$

- Logo encontramos um limite para a convergência uniforme considerando todas as **m** funções contidas no subespaço **finito** F na forma:

$$P(\sup_{f \in F} |R(f) - R_{emp}(f)| \geq \epsilon) \leq 2m \exp(-2n\epsilon^2)$$

- Observe que se o subespaço F for finito, **m** torna-se uma constante e a convergência uniforme ainda ocorre conforme n aumenta
- Prova-se, assim, que o princípio de Minimização do Risco Empírico sobre um subespaço finito de funções F é consistente
 - Mas como fica se o subespaço F for infinito?

- Finalmente obtemos um limite usando Chernoff para todo o conjunto de funções contidas no subespaço F na forma:

$$\sum_{i=1}^m P(|R(f_i) - R_{emp}(f_i)| \geq \epsilon) \leq 2m \exp(-2n\epsilon^2)$$

- Logo encontramos um limite para a convergência uniforme considerando todas as **m** funções contidas no subespaço **finito** F na forma:

Vapnik teve mais um desafio!
Como trabalhar com espaços infinitos de funções?

- Observe que se o subespaço F for finito, **m** torna-se uma constante e a convergência uniforme ainda ocorre conforme n aumenta
- Prova-se, assim, que o princípio de Minimização do Risco Empírico sobre um subespaço finito de funções F é consistente
 - Mas como fica se o subespaço F for infinito?

<p>Propriedades do subespaço F para Garantirmos Convergência Uniforme</p> <ul style="list-style-type: none"> • Antes, porém, Vapnik resolveu o problema de cálculo do Risco Esperado usando uma amostra fantasma (ghost sample) • Para isso utilizou o lema da Simetrização <ul style="list-style-type: none"> – O principal objetivo é transformar o termo abaixo apenas em função de amostras, pois não podemos computar o Risco Esperado ou Real! $\sup_{f \in \mathcal{F}} R(f) - R_{\text{emp}}(f) $ <ul style="list-style-type: none"> • Para isso Vapnik considerou uma amostra A e outra A' (fantasma) de exemplos rotulados. • Assumiu que essas amostras são independentes • Observe que a amostra A' não precisa ser obtida na prática, ela funciona como “e se obtivermos mais uma amostra”, ou seja, ela é o que se chama de ghost sample 	<p>Propriedades do subespaço F para Garantirmos Convergência Uniforme</p> <ul style="list-style-type: none"> • Antes, porém, Vapnik resolveu o problema de cálculo do Risco Esperado usando uma amostra fantasma (ghost sample) • Para isso utilizou o lema da Simetrização <ul style="list-style-type: none"> – O principal objetivo é transformar o termo abaixo apenas em função de amostras, pois não podemos computar o Risco Esperado ou Real! $\sup_{f \in \mathcal{F}} R(f) - R_{\text{emp}}(f) $ <ul style="list-style-type: none"> • Para isso Vapnik considerou uma amostra A e outra A' (fantasma) de exemplos rotulados. • Assumiu que essas amostras são independentes • Observe que a amostra A' não precisa ser obtida na prática, ela funciona como “e se obtivermos mais uma amostra”, ou seja, ela é o que se chama de ghost sample
<p>Propriedades do subespaço F para Garantirmos Convergência Uniforme</p> $P(\sup_{f \in \mathcal{F}} R(f) - R_{\text{emp}}(f) > \epsilon) \leq 2P(\sup_{f \in \mathcal{F}} R_{\text{emp}}(f) - R'_{\text{emp}}(f) > \epsilon/2)$	<p>Propriedades do subespaço F para Garantirmos Convergência Uniforme</p> <ul style="list-style-type: none"> • Tendo o ghost sample A', Vapnik chegou ao seguinte limite superior: $P(\sup_{f \in \mathcal{F}} R(f) - R_{\text{emp}}(f) > \epsilon) \leq 2P(\sup_{f \in \mathcal{F}} R_{\text{emp}}(f) - R'_{\text{emp}}(f) > \epsilon/2)$ <ul style="list-style-type: none"> • Cada amostra do lado direito tem tamanho n, por isso surge a constante 2 na formulação • Este lema é conhecido como lema da Simetrização <ul style="list-style-type: none"> • Apesar de não provarmos, é razoável que se os Riscos Empíricos em duas amostras diferentes e independentes, com n exemplos cada, sejam próximos <ul style="list-style-type: none"> – Então esses Riscos Empíricos também deveriam ser próximos ao Risco Esperado conforme n aumenta • Assim Vapnik trabalhou para retirar o termo R(f) que não podia ser calculado, pois somente temos acesso a amostras
<p>Propriedades do subespaço F para Garantirmos Convergência Uniforme</p> <ul style="list-style-type: none"> • Tendo o ghost sample A', Vapnik chegou ao seguinte limite superior: $P(\sup_{f \in \mathcal{F}} R(f) - R_{\text{emp}}(f) > \epsilon) \leq 2P(\sup_{f \in \mathcal{F}} R_{\text{emp}}(f) - R'_{\text{emp}}(f) > \epsilon/2)$ <ul style="list-style-type: none"> • Cada amostra do lado direito tem tamanho n, por isso surge a constante 2 na formulação • Este lema é conhecido como lema da Simetrização <ul style="list-style-type: none"> • Apesar de não provarmos, é razoável que se os Riscos Empíricos em duas amostras diferentes e independentes, com n exemplos cada, sejam próximos <ul style="list-style-type: none"> – Então esses Riscos Empíricos também deveriam ser próximos ao Risco Esperado conforme n aumenta • Assim Vapnik trabalhou para retirar o termo R(f) que não podia ser calculado, pois somente temos acesso a amostras 	<p>Propriedades do subespaço F para Garantirmos Convergência Uniforme</p> <ul style="list-style-type: none"> • Tendo o ghost sample A', Vapnik chegou ao seguinte limite superior: $P(\sup_{f \in \mathcal{F}} R(f) - R_{\text{emp}}(f) > \epsilon) \leq 2P(\sup_{f \in \mathcal{F}} R_{\text{emp}}(f) - R'_{\text{emp}}(f) > \epsilon/2)$ <ul style="list-style-type: none"> • Cada amostra do lado direito tem tamanho n, por isso surge a constante 2 na formulação • Este lema é conhecido como lema da Simetrização <ul style="list-style-type: none"> • Apesar de não provarmos, é razoável que se os Riscos Empíricos em duas amostras diferentes e independentes, com n exemplos cada, sejam próximos <ul style="list-style-type: none"> – Então esses Riscos Empíricos também deveriam ser próximos ao Risco Esperado conforme n aumenta • Assim Vapnik trabalhou para retirar o termo R(f) que não podia ser calculado, pois somente temos acesso a amostras

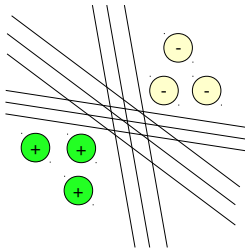
- Agora vejamos como o **lema da Simetrização** nos é útil:
 - Lembre-se que foi feita a prova de convergência do Princípio de Minimização do Risco Empírico para um subespaço F finito
 - Resta provarmos o mesmo para um subespaço F infinito
 - Para isso precisamos do **lema da Simetrização**
- É importante observar que se o subespaço F contém infinitas funções possíveis, a forma com que essas funções podem ser utilizadas para classificar um conjunto de treinamento com n exemplos é finita, por exemplo:
 - Assuma n exemplos na forma X_1, \dots, X_n
 - Assuma duas possíveis classes: $\{-1, +1\}$
 - Nesse caso, cada função f em F pode atuar no máximo de 2^n diferentes maneiras

- Até mesmo se o subespaço F contém infinitas funções, cada função f em F pode atuar no máximo de 2^n diferentes maneiras sobre um conjunto de treinamento de n exemplos
 - Isso significa que há infinitas funções similares que produzem a mesma classificação, por exemplo:

- Até mesmo se o subespaço F contém infinitas funções, cada função f em F pode atuar no máximo de 2^n diferentes maneiras sobre um conjunto de treinamento de n exemplos
 - Isso significa que há infinitas funções similares que produzem a mesma classificação, por exemplo:

Por exemplo, considere o subespaço F formado por todas retas

Infinitas retas produzem classificações equivalentes e, portanto, são consideradas como similares



$$\sup_{f \in \mathcal{F}} |R_{\text{emp}}(f) - R'_{\text{emp}}(f)|$$

$$\sup_{f \in \mathcal{F}} |R_{\text{emp}}(f) - R'_{\text{emp}}(f)|$$

- Portanto, o supremo executa apenas sobre um **conjunto finito de funções em F**:
 - Neste cenário, duas funções f e g, pertencentes a F, produzem o mesmo Risco Empírico $R_{\text{emp}}(f) = R_{\text{emp}}(g)$, desde que produzam classificações equivalentes
- Logo, considerando que temos duas amostras, teremos 2^n formas de classificar cada amostra e 2^{2n} formas de classificar as duas amostras em conjunto, dado que ambas contém n exemplos:
 - Ou seja, haverá no máximo 2^{2n} funções distintas considerando ambas amostras

- Portanto, o supremo executa apenas sobre um **conjunto finito de funções em F**:

$$\sup_{f \in \mathcal{F}} |R_{\text{emp}}(f) - R'_{\text{emp}}(f)|$$

- Neste cenário, duas funções f e g, pertencentes a F, produzem o mesmo Risco Empírico $R_{\text{emp}}(f) = R_{\text{emp}}(g)$, desde que produzam classificações equivalentes
- Logo, considerando que temos duas amostras, teremos 2^n formas de classificar cada amostra e 2^{2n} formas de classificar as duas amostras em conjunto, dado que ambas contém n exemplos:
 - Ou seja, haverá no máximo 2^{2n} funções distintas considerando ambas amostras

- Vapnik, então, utilizou os conceitos anteriores para derivar a primeira medida de capacidade para uma classe de funções F
 - Para isso, seja $Z_n = ((X_1, Y_1), \dots, (X_n, Y_n))$ um conjunto de exemplos de tamanho n
 - Seja $|F_{Z_n}|$ a cardinalidade do subespaço F para o conjunto de exemplos Z_n , ou seja, o número de funções que produzem classificações distintas para Z_n
 - Vamos denotar o número máximo de funções que podem ser distinguidas, ou seja, produzem classificações distintas por:

- Vapnik, então, utilizou os conceitos anteriores para derivar a primeira medida de capacidade para uma classe de funções F
 - Para isso, seja $Z_n = ((X_1, Y_1), \dots, (X_n, Y_n))$ um conjunto de exemplos de tamanho n
 - Seja $|F_{Z_n}|$ a cardinalidade do subespaço F para o conjunto de exemplos Z_n , ou seja, o número de funções que produzem classificações distintas para Z_n
 - Vamos denotar o número máximo de funções que podem ser distinguidas, ou seja, produzem classificações distintas por:

- Vapnik, então, utilizou os conceitos anteriores para derivar a primeira medida de capacidade para uma classe de funções F
 - Para isso, seja $Z_n = ((X_1, Y_1), \dots, (X_n, Y_n))$ um conjunto de exemplos de tamanho n
 - Seja $|F_{Z_n}|$ a cardinalidade do subespaço F para o conjunto de exemplos Z_n , ou seja, o número de funções que produzem classificações distintas para Z_n
 - Vamos denotar o número máximo de funções que podem ser distinguidas, ou seja, produzem classificações distintas por:

$$N(\mathcal{F}, n)$$

$$N(\mathcal{F}, n) = \max\{|F_{Z_n}| \mid X_1, \dots, X_n \in \mathcal{X}\}$$

<p>Propriedades do subespaço F para Garantirmos Convergência Uniforme</p> <ul style="list-style-type: none"> • Esse termo $N(\mathcal{F}, n)$ é denominado coeficiente de quebra ou coeficiente de shattering do subespaço F com respeito a um conjunto de n exemplos • Se $N(\mathcal{F}, n) = 2^n$, isso significa que existe ao menos uma amostra com n exemplos que pode ser classificada de todas as maneiras possíveis, assumindo duas possíveis classes $\{-1, +1\}$ <ul style="list-style-type: none"> • Neste caso dizemos que o subespaço F é capaz de “quebrar” (shatter) n pontos de todas formas possíveis • Observe que esse conceito considera a situação em que há ao menos uma amostra de tamanho n que pode ser quebrada de todas as maneiras possíveis (pois há um max na formulação do slide anterior) <ul style="list-style-type: none"> – Isso não implica que toda amostra pode ser quebrada (classificada) de todas maneiras possíveis! 	<p>Propriedades do subespaço F para Garantirmos Convergência Uniforme</p> <ul style="list-style-type: none"> • Esse termo $N(\mathcal{F}, n)$ é denominado coeficiente de quebra ou coeficiente de shattering do subespaço F com respeito a um conjunto de n exemplos • Se $N(\mathcal{F}, n) = 2^n$, isso significa que existe ao menos uma amostra com n exemplos que pode ser classificada de todas as maneiras possíveis, assumindo duas possíveis classes $\{-1, +1\}$ <ul style="list-style-type: none"> • Neste caso dizemos que o subespaço F é capaz de “quebrar” (shatter) n pontos de todas formas possíveis • Observe que esse conceito considera a situação em que há ao menos uma amostra de tamanho n que pode ser quebrada de todas as maneiras possíveis (pois há um max na formulação do slide anterior) <ul style="list-style-type: none"> – Isso não implica que toda amostra pode ser quebrada (classificada) de todas maneiras possíveis!
<p>Propriedades do subespaço F para Garantirmos Convergência Uniforme</p> <ul style="list-style-type: none"> • Esse termo $N(\mathcal{F}, n)$ é denominado coeficiente de quebra ou coeficiente de shattering do subespaço F com respeito a um conjunto de n exemplos • Se $N(\mathcal{F}, n) = 2^n$, isso significa que existe ao menos uma amostra com n exemplos que pode ser classificada de todas as maneiras possíveis, assumindo duas possíveis classes $\{-1, +1\}$ <ul style="list-style-type: none"> • Neste caso dizemos que o subespaço F é capaz de “quebrar” (shatter) n pontos de todas formas possíveis • Observe que esse conceito considera a situação em que há ao menos uma amostra de tamanho n que pode ser quebrada de todas as maneiras possíveis (pois há um max na formulação do slide anterior) <ul style="list-style-type: none"> – Isso não implica que toda amostra pode ser quebrada (classificada) de todas maneiras possíveis! 	<p>Propriedades do subespaço F para Garantirmos Convergência Uniforme</p>
<p>Propriedades do subespaço F para Garantirmos Convergência Uniforme</p> <ul style="list-style-type: none"> • Esse coeficiente de quebra é uma medida de capacidade da classe de funções F, i.e., ele permite medir o tamanho e/ou complexidade da classe de funções <ul style="list-style-type: none"> • Note que um subespaço F maior tende a possuir um maior coeficiente de quebra • Como Vapnik usou o coeficiente de shattering para produzir um limite superior para o lema da simetrização? $P(\sup_{f \in \mathcal{F}} R(f) - R_{\text{emp}}(f) > \epsilon) \leq 2P(\sup_{f \in \mathcal{F}} R_{\text{emp}}(f) - R'_{\text{emp}}(f) > \epsilon/2)$	<p>Propriedades do subespaço F para Garantirmos Convergência Uniforme</p>

Propriedades do subespaço F para Garantirmos Convergência Uniforme

- Como Vapnik usou esse coeficiente para produzir um limite superior para o **lema de Simetrização**?
 - Para isso considerou que há $2n$ exemplos, portanto há um conjunto Z_{2n} , em que os primeiros n exemplos são relativos à primeira amostra e os n seguintes à amostra fantasma (**ghost sample**)
 - Como primeiro passo Vapnik substituiu o supremo sobre F pelo supremo sobre $F_{Z_{2n}}$, sabendo que
 - Ou seja, o número máximo de funções que produzem classificações distintas para as duas amostras em conjunto é dado por 2^{2n}

Propriedades do subespaço F para Garantirmos Convergência Uniforme

- Como Vapnik usou esse coeficiente para produzir um limite superior para o **lema de Simetrização**?
 - Para isso considerou que há $2n$ exemplos, portanto há um conjunto Z_{2n} , em que os primeiros n exemplos são relativos à primeira amostra e os n seguintes à amostra fantasma (**ghost sample**)
 - Como primeiro passo Vapnik substituiu o supremo sobre F pelo supremo sobre $F_{Z_{2n}}$, sabendo que $N(\mathcal{F}, n) \leq 2^{2n}$
 - Ou seja, o número máximo de funções que produzem classificações distintas para as duas amostras em conjunto é dado por 2^{2n}

Propriedades do subespaço F para Garantirmos Convergência Uniforme

$$P(\sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)| > \epsilon) \leq 2P(\sup_{f \in \mathcal{F}} |R_{\text{emp}}(f) - R'_{\text{emp}}(f)| > \epsilon/2)$$

Propriedades do subespaço F para Garantirmos Convergência Uniforme

- Assim Vapnik partiu de:
$$P(\sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)| > \epsilon) \leq 2P(\sup_{f \in \mathcal{F}} |R_{\text{emp}}(f) - R'_{\text{emp}}(f)| > \epsilon/2)$$
- E obteve a igualdade:
$$2P(\sup_{f \in F} |R_{\text{emp}}(f) - R'_{\text{emp}}(f)| \geq \epsilon/2) = 2P(\sup_{f \in F_{Z_{2n}}} |R_{\text{emp}}(f) - R'_{\text{emp}}(f)| \geq \epsilon/2)$$
- Como $F_{Z_{2n}}$ contém no máximo $N(\mathcal{F}, 2n)$ funções distintas podemos chegar ao limite de Chernoff na forma:

Propriedades do subespaço F para Garantirmos Convergência Uniforme

- Assim Vapnik partiu de:
$$P(\sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)| > \epsilon) \leq 2P(\sup_{f \in \mathcal{F}} |R_{\text{emp}}(f) - R'_{\text{emp}}(f)| > \epsilon/2)$$
- E obteve a igualdade:
$$2P(\sup_{f \in F} |R_{\text{emp}}(f) - R'_{\text{emp}}(f)| \geq \epsilon/2) = 2P(\sup_{f \in F_{Z_{2n}}} |R_{\text{emp}}(f) - R'_{\text{emp}}(f)| \geq \epsilon/2)$$
- Como $F_{Z_{2n}}$ contém no máximo $N(\mathcal{F}, 2n)$ funções distintas podemos chegar ao limite de Chernoff na forma:
$$2P(\sup_{f \in F_{Z_{2n}}} |R_{\text{emp}}(f) - R'_{\text{emp}}(f)| \geq \epsilon/2) \leq 2N(\mathcal{F}, 2n) \exp(-n\epsilon^2/4)$$

Propriedades do subespaço F para Garantirmos Convergência Uniforme

- Assim Vapnik partiu de:
$$P(\sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)| > \epsilon) \leq 2P(\sup_{f \in \mathcal{F}} |R_{\text{emp}}(f) - R'_{\text{emp}}(f)| > \epsilon/2)$$
- E obteve
$$2P(\sup_{f \in F_{Z_{2n}}} |R_{\text{emp}}(f) - R'_{\text{emp}}(f)| \geq \epsilon/2) = 2P(\sup_{f \in F_{Z_{2n}}} |R_{\text{emp}}(f) - R'_{\text{emp}}(f)| \geq \epsilon/2)$$
- Como $F_{Z_{2n}}$ contém no máximo $N(\mathcal{F}, 2n)$ funções distintas podemos chegar ao limite de Chernoff na forma:
$$2P(\sup_{f \in F_{Z_{2n}}} |R_{\text{emp}}(f) - R'_{\text{emp}}(f)| \geq \epsilon/2) \leq \underline{2N(\mathcal{F}, 2n)} \exp(-n\epsilon^2/4)$$

- **Primeira conclusão:**
 - Considere que o coeficiente de quebra é consideravelmente menor que 2^{2n} , i.e., $\mathcal{N}(\mathcal{F}, 2n) \leq (2n)^k$ para alguma constante k
 - Isso indica que o coeficiente de quebra cresce polinomialmente, plugando isso no limite de Chernoff anteriormente obtido temos:
 - Vemos aqui que, conforme n tende ao infinito toda a expressão converge para zero
 - **Ou seja, o Princípio da Minimização do Risco Empírico é consistente com respeito a F caso o coeficiente de quebra tenha crescimento polinomial**

- **Primeira conclusão:**
 - Considere que o coeficiente de quebra é consideravelmente menor que 2^{2n} , i.e., $\mathcal{N}(\mathcal{F}, 2n) \leq (2n)^k$ para alguma constante k
 - Isso indica que o coeficiente de quebra cresce polinomialmente, plugando isso no limite de Chernoff anteriormente obtido temos:
$$2\mathcal{N}(\mathcal{F}, 2n) \exp(-n\varepsilon^2/4) = 2 \cdot (2n)^k \cdot \exp(-n\varepsilon^2/4) = 2 \exp(k \cdot \log(2n) - n\varepsilon^2/4)$$
 - Vemos aqui que, conforme n tende ao infinito toda a expressão converge para zero
 - **Ou seja, o Princípio da Minimização do Risco Empírico é consistente com respeito a F caso o coeficiente de quebra tenha crescimento polinomial**
- **Primeira conclusão:**
 - Considere que o coeficiente de quebra é consideravelmente menor que 2^{2n} , i.e., $\mathcal{N}(\mathcal{F}, 2n) \leq (2n)^k$ para alguma constante k
 - Isso indica que o coeficiente de quebra cresce polinomialmente, plugando isso no limite de Chernoff anteriormente obtido temos:
$$2\mathcal{N}(\mathcal{F}, 2n) \exp(-n\varepsilon^2/4) = 2 \cdot (2n)^k \cdot \exp(-n\varepsilon^2/4) = 2 \exp(k \cdot \log(2n) - n\varepsilon^2/4)$$
 - Vemos aqui que, conforme n tende ao infinito toda a expressão converge para zero
 - **Ou seja, o Princípio da Minimização do Risco Empírico é consistente com respeito a F caso o coeficiente de quebra tenha crescimento polinomial**

- **Primeira conclusão:**
 - Considere que o coeficiente de quebra é consideravelmente menor que 2^{2n} , i.e., $\mathcal{N}(\mathcal{F}, 2n) \leq (2n)^k$ para alguma constante k
 - Isso indica que o coeficiente de quebra cresce polinomialmente, plugando isso no limite de Chernoff anteriormente obtido temos:
$$2\mathcal{N}(\mathcal{F}, 2n) \exp(-n\varepsilon^2/4) = 2 \cdot (2n)^k \cdot \exp(-n\varepsilon^2/4) = 2 \exp(k \cdot \log(2n) - n\varepsilon^2/4)$$
 - Vemos aqui que, conforme n tende ao infinito toda a expressão converge para zero
 - **Ou seja, o Princípio da Minimização do Risco Empírico é consistente com respeito a F caso o coeficiente de quebra tenha crescimento polinomial**

<p>Propriedades do subespaço F para Garantirmos Convergência Uniforme</p> <ul style="list-style-type: none"> • Segunda conclusão: <ul style="list-style-type: none"> • Considere o caso em que utilizamos F_{all}, i.e., consideramos o espaço com todas as possíveis funções <ul style="list-style-type: none"> – É claro que nessa situação poderemos classificar o conjunto de exemplos de todas as maneiras possíveis ou seja: – Assim, ao substituírmos no limite de Chernoff obtemos: – Como epsilon assume um valor muito pequeno, logo vemos que conforme n aumenta essa expressão não tende a zero 	<p>Propriedades do subespaço F para Garantirmos Convergência Uniforme</p> <ul style="list-style-type: none"> • Segunda conclusão: <ul style="list-style-type: none"> • Considere o caso em que utilizamos F_{all}, i.e., consideramos o espaço com todas as possíveis funções <ul style="list-style-type: none"> – É claro que nessa situação poderemos classificar o conjunto de exemplos de todas as maneiras possíveis ou seja: – Assim, ao substituírmos no limite de Chernoff obtemos: – Como epsilon assume um valor muito pequeno, logo vemos que conforme n aumenta essa expressão não tende a zero
<p>Propriedades do subespaço F para Garantirmos Convergência Uniforme</p> <ul style="list-style-type: none"> • Segunda conclusão: <ul style="list-style-type: none"> • Considere o caso em que utilizamos F_{all}, i.e., consideramos o espaço com todas as possíveis funções <ul style="list-style-type: none"> – É claro que nessa situação poderemos classificar o conjunto de exemplos de todas as maneiras possíveis ou seja: – Assim, ao substituírmos no limite de Chernoff obtemos: – Como epsilon assume um valor muito pequeno, logo vemos que conforme n aumenta essa expressão não tende a zero 	<p>Propriedades do subespaço F para Garantirmos Convergência Uniforme</p> <ul style="list-style-type: none"> • Segunda conclusão: <ul style="list-style-type: none"> • Considere o caso em que utilizamos F_{all}, i.e., consideramos o espaço com todas as possíveis funções <ul style="list-style-type: none"> – É claro que nessa situação poderemos classificar o conjunto de exemplos de todas as maneiras possíveis ou seja: – Assim, ao substituírmos no limite de Chernoff obtemos: – Como epsilon assume um valor muito pequeno, logo vemos que conforme n aumenta essa expressão não tende a zero
<p>Propriedades do subespaço F para Garantirmos Convergência Uniforme</p> <ul style="list-style-type: none"> • Segunda conclusão: <ul style="list-style-type: none"> • Considere o caso em que utilizamos F_{all}, i.e., consideramos o espaço com todas as possíveis funções <ul style="list-style-type: none"> – É claro que nessa situação poderemos classificar o conjunto de exemplos de todas as maneiras possíveis ou seja: – Assim, ao substituírmos no limite de Chernoff obtemos: – Como epsilon assume um valor muito pequeno, logo vemos que conforme n aumenta essa expressão não tende a zero 	<p>Propriedades do subespaço F para Garantirmos Convergência Uniforme</p> <ul style="list-style-type: none"> • Segunda conclusão: <ul style="list-style-type: none"> • Logo não podemos concluir consistência do Princípio de Minimização do Risco Empírico para F_{all} <ul style="list-style-type: none"> – Por outro lado ainda não podemos concluir que o princípio é inconsistente para F_{all} – Isso se deve ao fato de que o lado direito da desigualdade abaixo provê uma condição suficiente para consistência e não uma condição necessária – No entanto, mais tarde Vapnik e Chervonenkis provaram que a condição a seguir é necessária para garantir consistência do Princípio de Minimização do Risco Empírico:

Propriedades do subespaço F para Garantirmos Convergência Uniforme	Propriedades do subespaço F para Garantirmos Convergência Uniforme
<ul style="list-style-type: none"> Segunda conclusão: <ul style="list-style-type: none"> Logo não podemos concluir consistência do Princípio de Minimização do Risco Empírico para F_{all} <ul style="list-style-type: none"> Por outro lado ainda não podemos concluir que o princípio é inconsistente para F_{all} Isso se deve ao fato de que o lado direito da desigualdade abaixo provê uma condição suficiente para consistência e não uma condição necessária $P(\sup_{f \in \mathcal{F}} R(f) - R_{emp}(f) > \epsilon) \leq 2N(\mathcal{F}, 2n) \exp(-n\epsilon^2/4)$ <ul style="list-style-type: none"> No entanto, mais tarde Vapnik e Chervonenkis provaram que a condição a seguir é necessária para garantir consistência do Princípio de Minimização do Risco Empírico: 	<ul style="list-style-type: none"> Segunda conclusão: <ul style="list-style-type: none"> Logo não podemos concluir consistência do Princípio de Minimização do Risco Empírico para F_{all} <ul style="list-style-type: none"> Por outro lado ainda não podemos concluir que o princípio é inconsistente para F_{all} Isso se deve ao fato de que o lado direito da desigualdade abaixo provê uma condição suficiente para consistência e não uma condição necessária $P(\sup_{f \in \mathcal{F}} R(f) - R_{emp}(f) > \epsilon) \leq 2N(\mathcal{F}, 2n) \exp(-n\epsilon^2/4)$ <ul style="list-style-type: none"> No entanto, mais tarde Vapnik e Chervonenkis provaram que a condição a seguir é necessária para garantir consistência do Princípio de Minimização do Risco Empírico: $\frac{\log \mathcal{N}(F, n)}{n} \rightarrow 0$

Minimização do Risco Empírico	Limites de Generalização
<ul style="list-style-type: none"> Assim: <ul style="list-style-type: none"> Se $N(F, n)$ é polinomial, então a condição tende a zero Se considerarmos um espaço com todas funções possíveis, ou seja, F_{all}, teremos $N(F, n) = 2^n$, logo $\frac{\log \mathcal{N}(F, n)}{n} = \frac{\log 2^n}{n} = \frac{n}{n} = 1$ <ul style="list-style-type: none"> E, assim, a minimização de Risco Empírico não é consistente para F_{all} 	$P(\sup_{f \in \mathcal{F}} R(f) - R_{emp}(f) > \epsilon) \leq 2N(\mathcal{F}, 2n) \exp(-n\epsilon^2/4)$

Limites de Generalização	Limites de Generalização
<ul style="list-style-type: none"> Podemos escrever a desigualdade abaixo de outra maneira: $P(\sup_{f \in \mathcal{F}} R(f) - R_{emp}(f) > \epsilon) \leq 2N(\mathcal{F}, 2n) \exp(-n\epsilon^2/4)$ <ul style="list-style-type: none"> Ao invés de fixar epsilon e computar a probabilidade do Risco Empírico desviar do Risco Esperado, podemos especificar a probabilidade delta para a qual o limite seja válido: <ul style="list-style-type: none"> Para isso igualamos o lado direito a um delta maior que zero na forma: E agora resolvemos para epsilon... 	<ul style="list-style-type: none"> Podemos escrever a desigualdade abaixo de outra maneira: $P(\sup_{f \in \mathcal{F}} R(f) - R_{emp}(f) > \epsilon) \leq 2N(\mathcal{F}, 2n) \exp(-n\epsilon^2/4)$ <ul style="list-style-type: none"> Ao invés de fixar epsilon e computar a probabilidade do Risco Empírico desviar do Risco Esperado, podemos especificar a probabilidade delta para a qual o limite seja válido: <ul style="list-style-type: none"> Para isso igualamos o lado direito a um delta maior que zero na forma: E agora resolvemos para epsilon... $P(\sup_{f \in \mathcal{F}} R(f) - R_{emp}(f) > \epsilon) \leq \delta$

- Podemos escrever a desigualdade abaixo de outra maneira:

$$P(\sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)| > \epsilon) \leq 2\mathcal{N}(\mathcal{F}, 2n) \exp(-n\epsilon^2/4)$$

- Ao invés de fixar epsilon e computar a probabilidade do Risco Empírico desviar do Risco Esperado, podemos especificar a probabilidade delta para a qual o limite seja válido:

- Para isso igualamos o lado direito a um delta maior que zero na forma:

$$P(\sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)| > \epsilon) \leq \delta$$

- E agora resolvemos para epsilon...

$$2\mathcal{N}(F, 2n) \exp(-n\epsilon^2/4) = \delta \quad \text{para } \delta > 0$$

- Resolvendo para epsilon:

$$2\mathcal{N}(F, 2n) \exp(-n\epsilon^2/4) = \delta$$

$$\exp(-n\epsilon^2/4) = \frac{\delta}{2\mathcal{N}(F, 2n)}$$

$$\log(\exp(-n\epsilon^2/4)) = \log(\delta) - \log(2\mathcal{N}(F, 2n))$$

$$-n\epsilon^2/4 = \log(\delta) - \log(2\mathcal{N}(F, 2n))$$

$$\epsilon^2 = -\frac{4}{n} (\log(\delta) - \log(2\mathcal{N}(F, 2n)))$$

$$\epsilon = \sqrt{\frac{4}{n} (\log(2\mathcal{N}(F, 2n)) - \log(\delta))}$$

$$P(\sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)| > \epsilon) \leq 2\mathcal{N}(\mathcal{F}, 2n) \exp(-n\epsilon^2/4)$$

$$P(\sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)| > \epsilon) \leq \delta$$

- Como substituímos delta conforme apresentado abaixo:

$$P(\sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)| > \epsilon) \leq 2\mathcal{N}(\mathcal{F}, 2n) \exp(-n\epsilon^2/4)$$

$$P(\sup_{f \in \mathcal{F}} |R(f) - R_{\text{emp}}(f)| > \epsilon) \leq \delta$$

- Podemos considerar, então que o desvio do Risco Esperado para o Risco Empírico é no máximo igual a epsilon, o que nos permite obter:

$$R(f) \leq R_{\text{emp}}(f) + \sqrt{\frac{4}{n} (\log(2\mathcal{N}(\mathcal{F}, n)) - \log(\delta))}$$

- Agora fica óbvio de onde Vapnik e Chervonenkis obtiveram:

$$\frac{\log \mathcal{N}(F, n)}{n} \rightarrow 0$$

- Pois chegamos a:

$$R(f) \leq R_{\text{emp}}(f) + \sqrt{\frac{4}{n} (\log(2\mathcal{N}(\mathcal{F}, n)) - \log(\delta))}$$

$$R(f) \leq R_{\text{emp}}(f) + \sqrt{\frac{4}{n} (\log(2\mathcal{N}(\mathcal{F}, n)) - \log(\delta))}$$

Limites de Generalização

- Vamos analisar intuitivamente esse limite:

$$R(f) \leq R_{\text{emp}}(f) + \sqrt{\frac{4}{n} (\log(2N(\mathcal{F}, n)) - \log(\delta))}$$

- Se o Risco Empírico for pequeno e o termo na raiz também, então há grande probabilidade do Risco Esperado ser baixo
 - Se:
 - Usarmos um subespaço F pequeno, ou seja, com poucas funções que produzem classificações distintas
 - E esse subespaço ainda for capaz de representar o conjunto de exemplos de treinamento
 - Então podemos concluir que há grande probabilidade de aprendermos um conceito

Limites de Generalização

- Vamos analisar intuitivamente esse limite:

$$R(f) \leq R_{\text{emp}}(f) + \sqrt{\frac{4}{n} (\log(2N(\mathcal{F}, n)) - \log(\delta))}$$

- Se o Risco Empírico for pequeno e o termo na raiz também, então há grande probabilidade do Risco Esperado ser baixo
 - Se:
 - Usarmos um subespaço F pequeno, ou seja, com poucas funções que produzem classificações distintas
 - E esse subespaço ainda for capaz de representar o conjunto de exemplos de treinamento
 - Então podemos concluir que há grande probabilidade de aprendermos um conceito

Limites de Generalização

- Vamos analisar intuitivamente esse limite:

$$R(f) \leq R_{\text{emp}}(f) + \sqrt{\frac{4}{n} (\log(2N(\mathcal{F}, n)) - \log(\delta))}$$

- Se o Risco Empírico for pequeno e o termo na raiz também, então há grande probabilidade do Risco Esperado ser baixo
 - Se:
 - Usarmos um subespaço F pequeno, ou seja, com poucas funções que produzem classificações distintas
 - E esse subespaço ainda for capaz de representar o conjunto de exemplos de treinamento
 - Então podemos concluir que há grande probabilidade de aprendermos um conceito

Limites de Generalização

- Vamos analisar intuitivamente esse limite:

$$R(f) \leq R_{\text{emp}}(f) + \sqrt{\frac{4}{n} (\log(2N(\mathcal{F}, n)) - \log(\delta))}$$

- Ou seja, há grande probabilidade de aprendizado quando:
 - Há viés, i.e., definimos um subespaço F
 - Esse subespaço é capaz de representar os exemplos de treinamento amostrados

Limites de Generalização

$$R(f) \leq R_{\text{emp}}(f) + \sqrt{\frac{4}{n} (\log(2N(\mathcal{F}, n)) - \log(\delta))}$$

- Por outro lado o que ocorre se nosso problema é muito complexo?**

$$R(f) \leq R_{\text{emp}}(f) + \sqrt{\frac{4}{n} (\log(2N(\mathcal{F}, n)) - \log(\delta))}$$

- Precisamos de um subespaço F maior para explicar os exemplos de treinamento amostrados
 - O problema é que conforme aumentamos o subespaço F , podemos chegar ao espaço F_{all} capaz de representar qualquer classificação
 - No entanto, esse espaço muito grande faz com que o que não nos permite definir o limite superior, assim não podemos dizer que o Risco Empírico é um bom estimador para o Risco Esperado

- Por outro lado o que ocorre se nosso problema é muito complexo?

$$R(f) \leq R_{\text{emp}}(f) + \sqrt{\frac{4}{n} (\log(2N(\mathcal{F}, n)) - \log(\delta))}$$

- Precisamos de um subespaço F maior para explicar os exemplos de treinamento amostrados
 - O problema é que conforme aumentamos o subespaço F , podemos chegar ao espaço F_{all} capaz de representar qualquer classificação
 - No entanto, esse espaço muito grande faz com que o que não nos permite definir o limite superior, assim não podemos dizer que o Risco Empírico é um bom estimador para o Risco Esperado

- Por outro lado o que ocorre se nosso problema é muito complexo?

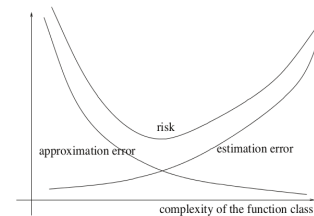
$$R(f) \leq R_{\text{emp}}(f) + \sqrt{\frac{4}{n} (\log(2N(\mathcal{F}, n)) - \log(\delta))}$$

- Precisamos de um subespaço F maior para explicar os exemplos de treinamento amostrados
 - O problema é que conforme aumentamos o subespaço F , podemos chegar ao espaço F_{all} capaz de representar qualquer classificação
 - No entanto, esse espaço muito grande faz com que $N(F, 2n) \rightarrow 2^{2n}$ o que não nos permite definir o limite superior, assim não podemos dizer que o Risco Empírico é um bom estimador para o Risco Esperado

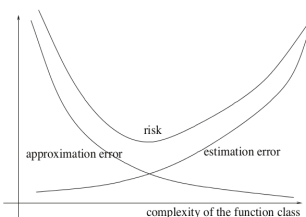
- Por outro lado o que ocorre se nosso problema é muito complexo?

$$R(f) \leq R_{\text{emp}}(f) + \sqrt{\frac{4}{n} (\log(2N(\mathcal{F}, n)) - \log(\delta))}$$

- No entanto, mesmo para um problema muito complexo, poderíamos definir um viés, ou seja, um subespaço F a ser considerado
 - Assim o Princípio de Minimização do Risco Empírico seria consistente

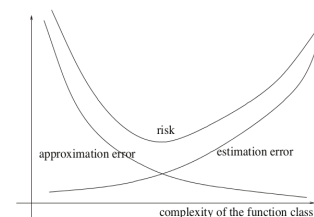


- Em resumo:
 - A definição de um viés é essencial para que o Princípio de Minimização do Risco Empírico seja consistente
 - O que nos leva a garantias de aprendizado



- Observe a importância em avaliar diferentes técnicas (**com diferentes vieses**) para abordar um mesmo problema:
 - Muitos pesquisadores trabalham com ensembles
- A formulação do slide anterior é muito similar à regularização de Tikonov, comumente utilizada na área de Processamento de Sinais

- Em resumo:
 - A definição de um viés é essencial para que o Princípio de Minimização do Risco Empírico seja consistente
 - O que nos leva a garantias de aprendizado



- Observe a importância em avaliar diferentes técnicas (**com diferentes vieses**) para abordar um mesmo problema:
 - Muitos pesquisadores trabalham com ensembles
- A formulação do slide anterior é muito similar à regularização de Tikonov, comumente utilizada na área de Processamento de Sinais

Dimensão VC	Dimensão VC
<ul style="list-style-type: none"> O coeficiente de quebra motivou Vapnik e Chervonenkis a propor outra medida de capacidade para um conjunto de funções 	


Dimensão VC	Dimensão VC
<ul style="list-style-type: none"> Vapnik e Chervonenkis propuseram uma medida para caracterizar o crescimento do coeficiente de quebra: <ul style="list-style-type: none"> Dimensão VC (Vapnik-Chervonenkis) Vapnik e Chervonenkis assumem que: <ul style="list-style-type: none"> Uma amostra de n elementos Z_n é “quebrada” por uma classe de funções F se tal classe pode realizar qualquer classificação em dada amostra, i.e., a cardinalidade de $F_{Z_n} = 2^n$ Assim, a Dimensão VC de F é definida como o maior número n tal que existe uma amostra de tamanho n que pode ser “quebrada” por F Se o máximo não existe, a dimensão VC é definida como infinita 	<ul style="list-style-type: none"> Vapnik e Chervonenkis propuseram uma medida para caracterizar o crescimento do coeficiente de quebra: <ul style="list-style-type: none"> Dimensão VC (Vapnik-Chervonenkis) Vapnik e Chervonenkis assumem que: <ul style="list-style-type: none"> Uma amostra de n elementos Z_n é “quebrada” por uma classe de funções F se tal classe pode realizar qualquer classificação em dada amostra, i.e., a cardinalidade de $F_{Z_n} = 2^n$ Assim, a Dimensão VC de F é definida como o maior número n tal que existe uma amostra de tamanho n que pode ser “quebrada” por F <div> $VC(F) = \max(n \in \mathbb{Z}^+ \mid F_{Z_n} = 2^n \text{ para algum } Z_n)$ </div> <ul style="list-style-type: none"> Se o máximo não existe, a dimensão VC é definida como infinita

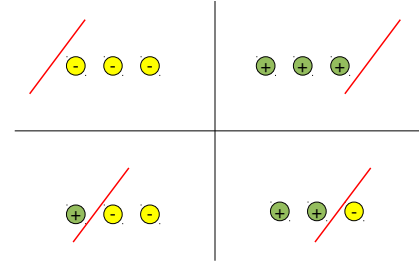
Dimensão VC	Dimensão VC
<ul style="list-style-type: none"> Vapnik e Chervonenkis propuseram uma medida para caracterizar o crescimento do coeficiente de quebra: <ul style="list-style-type: none"> Dimensão VC (Vapnik-Chervonenkis) Vapnik e Chervonenkis assumem que: <ul style="list-style-type: none"> Uma amostra de n elementos Z_n é “quebrada” por uma classe de funções F se tal classe pode realizar qualquer classificação em dada amostra, i.e., a cardinalidade de $F_{Z_n} = 2^n$ Assim, a Dimensão VC de F é definida como o maior número n tal que existe uma amostra de tamanho n que pode ser “quebrada” por F <div> $VC(F) = \max(n \in \mathbb{Z}^+ \mid F_{Z_n} = 2^n \text{ para algum } Z_n)$ </div> <ul style="list-style-type: none"> Se o máximo não existe, a dimensão VC é definida como infinita 	

Dimensão VC

- Assim, uma vez que sabemos que a dimensão VC de uma classe de funções em F é finita:
 - Concluimos que o coeficiente de shattering cresce polinomialmente conforme o tamanho da amostra aumenta
 - Isso implica na consistência do Princípio de Minimização do Risco Empírico
 - O que implica, em resumo, em aprendizado!
- Se a dimensão VC é infinita então existe alguma amostra que pode ser “quebrada” por F de todas as maneiras possíveis, i.e., em 2^n maneiras
 - Nesse caso, conforme visto anteriormente, o Princípio de Minimização do Risco Empírico não é consistente
 - Logo, não há garantias de aprendizado

Dimensão VC

- Por exemplo, sejam três pontos
 
- Assumindo todas retas podemos dividir em:



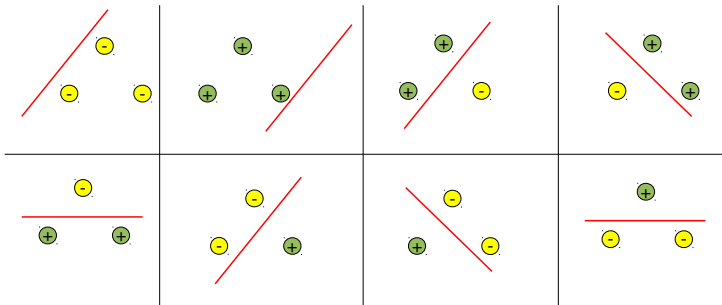
- Nesta amostra pudemos “quebrar” ou classificar de 4 formas diferentes
- Mas será que há outra amostra de 3 pontos que poderíamos “quebrar” de mais maneiras?

Dimensão VC

- Mas esses três pontos podem, ainda ser organizados de maneira distinta:



- Assumindo todas retas podemos dividir em:




- Perceba que pudemos “quebrar” esta amostra de todas as 2^3 maneiras possíveis, o que nos leva ao fato de que a dimensão VC é pelo menos igual a 3
 - Pois há ao menos uma amostra de 3 exemplos que pode ser “quebrada” de todas as maneiras possíveis nesse espaço bidimensional

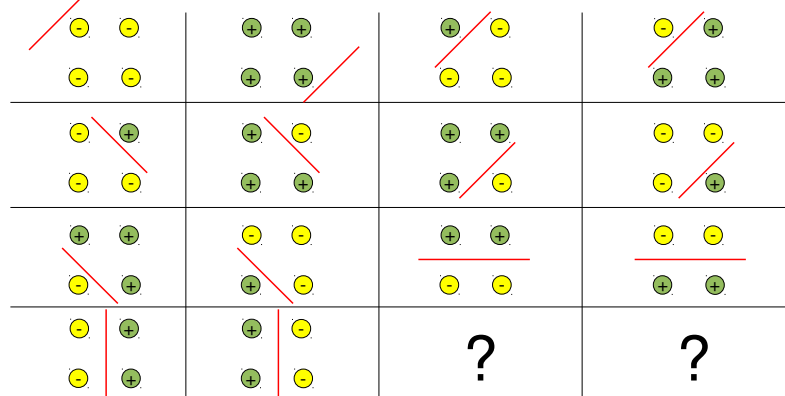
Dimensão VC

Dimensão VC

- Assim, para um problema definido em \mathbb{R}^2 e com classificadores baseados em retas, podemos “quebrar” exemplos em pelo menos:
 - $2^3 = 8$ formas, então sabemos que a dimensão VC é igual ou ainda maior que 3, i.e., $VC \geq 3$
 - Portanto, ainda não conhecemos a dimensão VC, assim se diz que seu valor mínimo é 3
- Como a dimensão VC é dada pelo valor máximo para o problema, basta ter um caso em que é possível “quebrar” os pontos de todas as formas possíveis para chegar na dimensão VC
 - Mas, que tal tentarmos “quebrar” mais pontos nesse mesmo plano \mathbb{R}^2 ainda com classificadores baseados em retas?
 - Talvez a dimensão VC seja ainda maior!!!

Dimensão VC

- Seja:
 
- Assumindo todas retas podemos dividir em:



- Há formas de conduzir provas para chegar na dimensão VC sem precisar verificar toda possível organização de amostras, o que é muito complexo!
- Para provar isso necessitamos provar que não há nenhum conjunto de quatro pontos que possa ser dividido de todas maneiras
 - **Essa prova é difícil**
- No entanto, já há uma prova que hiperplanos lineares de $n-1$ dimensões utilizados para dividir um espaço R^n apresentam dimensão VC = $n+1$

- Há formas de conduzir provas para chegar na dimensão VC sem precisar verificar toda possível organização de amostras, o que é muito complexo!
- Para provar isso necessitamos provar que não há nenhum conjunto de quatro pontos que possa ser dividido de todas maneiras
 - **Essa prova é difícil**
- No entanto, já há uma prova que hiperplanos lineares de $n-1$ dimensões utilizados para dividir um espaço R^n apresentam dimensão VC = $n+1$

- Há formas de conduzir provas para chegar na dimensão VC sem precisar verificar toda possível organização de amostras, o que é muito complexo!
- Para provar isso necessitamos provar que não há nenhum conjunto de quatro pontos que possa ser dividido de todas maneiras
 - **Essa prova é difícil**
- No entanto, já há uma prova que hiperplanos lineares de $n-1$ dimensões utilizados para dividir um espaço R^n apresentam dimensão VC = $n+1$

Observe a conclusão:

A partir da dimensão VC do conjunto de funções definidas pelo viés de um Algoritmo de Classificação, podemos concluir se o Princípio de Minimização do Risco Empírico é consistente e, portanto, concluir se ocorre aprendizado!

Lembre-se: O oposto não é garantido, i.e., pode ocorrer aprendizado mesmo Quando o Princípio é inconsistente.

- Finalmente veremos mais uma medida de capacidade para uma classe de funções
 - Considere o caso em que utilizaremos retas para separar pontos em um espaço R^2
 - Considere um conjunto de pontos previamente rotulados e que podem ser perfeitamente separados
 - Define-se a margem do classificador f como a menor distância existente entre qualquer ponto do conjunto de treinamento e a reta

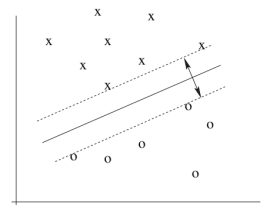


Figura obtida de Luxburg and Scholkopf, Statistical Learning Theory: Models, Concepts, and Results

Limites para Margens

- Finalmente veremos mais uma medida de capacidade para uma classe de funções
 - Considere o caso em que utilizaremos retas para separar pontos em um espaço \mathbb{R}^2
 - Considere um conjunto de pontos previamente rotulados e que podem ser perfeitamente separados
 - Define-se a margem do classificador f como a menor distância existente entre qualquer ponto do conjunto de treinamento e a reta

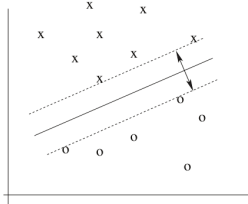


Figura obtida de Luxburg and Scholkopf, Statistical Learning Theory: Models, Concepts, and Results

Limites para Margens

- Finalmente veremos mais uma medida de capacidade para uma classe de funções
 - Considere o caso em que utilizaremos retas para separar pontos em um espaço \mathbb{R}^2
 - Considere um conjunto de pontos previamente rotulados e que podem ser perfeitamente separados
 - Define-se a margem do classificador f como a menor distância existente entre qualquer ponto do conjunto de treinamento e a reta

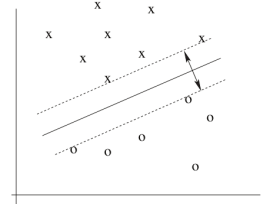


Figura obtida de Luxburg and Scholkopf, Statistical Learning Theory: Models, Concepts, and Results

Limites para Margens

$$VC(F_p) \leq \min_{\mathbb{R}^d} \left(d, \frac{4R^2}{p^2} \right) + 1$$

Limites para Margens

- Há uma prova de que a Dimensão VC de uma classe de funções lineares F_p que apresentam margem de pelo menos p pode ser limitada pela taxa do raio R da menor esfera ao redor dos pontos de dados com margem p :

$$VC(F_p) \leq \min \left(d, \frac{4R^2}{p^2} \right) + 1$$

- Em que d é a dimensão do espaço \mathbb{R}^d
- Consequentemente, conforme provado por Vapnik, quanto maior a margem p , menor a dimensão VC
 - Logo a margem pode ser utilizada como medida de capacidade para uma classe de funções → **O que de fato motivou SVM (Support Vector Machines)**

Limites para Margens

- Vapnik ainda provou o limite de margem utilizando como base o Princípio de Minimização do Risco Empírico

$$R(f) \leq \nu(f) + \sqrt{\frac{c}{n} \left(\frac{R^2}{\rho^2} \log(n)^2 + \log(1/\delta) \right)}$$

- Em que $\nu(f)$ é uma fração dos exemplos de treinamento que apresentam margem ao menos igual a p
- Assim a maximização de margem dá consistência para o Princípio de Minimização do Risco Empírico

Conclusões

Conclusões	Conclusões
<ul style="list-style-type: none"> A Teoria do Aprendizado Estatístico permite verificar sob quais condições ocorre aprendizado Observe que: <ul style="list-style-type: none"> Sem viés, o Princípio de Minimização do Risco Empírico não é consistente <ul style="list-style-type: none"> Ou seja, precisamos de algum viés para aprendermos Calculando a dimensão VC para o conjunto de funções considerado, verificamos a consistência do Princípio de Minimização do Risco Empírico Podemos pesquisar na área de algoritmos que adaptam seus vieses <ul style="list-style-type: none"> Isso permitiria explorarmos melhor o espaço e, talvez, chegarmos a um subespaço F que contenha f_{Bayes} Apesar de haver vários estudos prévios a Vladimir Vapnik, este pesquisador em conjunto com outros, dentre eles: Alexey Chervonenkis, desenvolveram grande parte dos conceitos apresentados 	<ul style="list-style-type: none"> A Teoria do Aprendizado Estatístico permite verificar sob quais condições ocorre aprendizado Observe que: <ul style="list-style-type: none"> Sem viés, o Princípio de Minimização do Risco Empírico não é consistente <ul style="list-style-type: none"> Ou seja, precisamos de algum viés para aprendermos Calculando a dimensão VC para o conjunto de funções considerado, verificamos a consistência do Princípio de Minimização do Risco Empírico Podemos pesquisar na área de algoritmos que adaptam seus vieses <ul style="list-style-type: none"> Isso permitiria explorarmos melhor o espaço e, talvez, chegarmos a um subespaço F que contenha f_{Bayes} Apesar de haver vários estudos prévios a Vladimir Vapnik, este pesquisador em conjunto com outros, dentre eles: Alexey Chervonenkis, desenvolveram grande parte dos conceitos apresentados

Conclusões	Conclusões
<ul style="list-style-type: none"> A Teoria do Aprendizado Estatístico permite verificar sob quais condições ocorre aprendizado Observe que: <ul style="list-style-type: none"> Sem viés, o Princípio de Minimização do Risco Empírico não é consistente <ul style="list-style-type: none"> Ou seja, precisamos de algum viés para aprendermos Calculando a dimensão VC para o conjunto de funções considerado, verificamos a consistência do Princípio de Minimização do Risco Empírico Podemos pesquisar na área de algoritmos que adaptam seus vieses <ul style="list-style-type: none"> Isso permitiria explorarmos melhor o espaço e, talvez, chegarmos a um subespaço F que contenha f_{Bayes} Apesar de haver vários estudos prévios a Vladimir Vapnik, este pesquisador em conjunto com outros, dentre eles: Alexey Chervonenkis, desenvolveram grande parte dos conceitos apresentados 	<ul style="list-style-type: none"> A Teoria do Aprendizado Estatístico permite verificar sob quais condições ocorre aprendizado Observe que: <ul style="list-style-type: none"> Sem viés, o Princípio de Minimização do Risco Empírico não é consistente <ul style="list-style-type: none"> Ou seja, precisamos de algum viés para aprendermos Calculando a dimensão VC para o conjunto de funções considerado, verificamos a consistência do Princípio de Minimização do Risco Empírico Podemos pesquisar na área de algoritmos que adaptam seus vieses <ul style="list-style-type: none"> Isso permitiria explorarmos melhor o espaço e, talvez, chegarmos a um subespaço F que contenha f_{Bayes} Apesar de haver vários estudos prévios a Vladimir Vapnik, este pesquisador em conjunto com outros, dentre eles: Alexey Chervonenkis, desenvolveram grande parte dos conceitos apresentados

Conclusões	Referências
<ul style="list-style-type: none"> A Teoria do Aprendizado Estatístico permite verificar sob quais condições ocorre aprendizado Observe que: <ul style="list-style-type: none"> Sem viés, o Princípio de Minimização do Risco Empírico não é consistente <ul style="list-style-type: none"> Ou seja, precisamos de algum viés para aprendermos Calculando a dimensão VC para o conjunto de funções considerado, verificamos a consistência do Princípio de Minimização do Risco Empírico Podemos pesquisar na área de algoritmos que adaptam seus vieses <ul style="list-style-type: none"> Isso permitiria explorarmos melhor o espaço e, talvez, chegarmos a um subespaço F que contenha f_{Bayes} Apesar de haver vários estudos prévios a Vladimir Vapnik, este pesquisador em conjunto com outros, dentre eles: Alexey Chervonenkis, desenvolveram grande parte dos conceitos apresentados 	<ul style="list-style-type: none"> Luxburg and Scholkopf, Statistical Learning Theory: Models, Concepts, and Results. Handbook of the History of Logic. Volume 10: Inductive Logic. Volume Editors: Dov M. Gabbay, Stephan Hartmann and John Woods, Elsevier, 2009

Avaliando Algoritmos de Aprendizado segundo a Teoria do Aprendizado Estatístico	Outros Conceitos Relevantes
<ul style="list-style-type: none"> • Árvores de Decisão <ul style="list-style-type: none"> • ID3 • C4.5 • Redes Neurais <ul style="list-style-type: none"> • Perceptron • Multilayer Perceptron • Aprendizado Baseado em Instâncias <ul style="list-style-type: none"> • K-Nearest Neighbors • Máquinas de Vetores de Suporte 	<ul style="list-style-type: none"> • Complexidade de Rademacher • Probably Approximately Correct (PAC) Learning • Luckiness Approach • No Free Lunch Theorem • Minimum Description Length Principle

Últimas Formalizações	Outros Assuntos em Preparação
<ul style="list-style-type: none"> • Bousquet & Eliseeff (2002), Stability and Generalization. Journal of Machine Learning Research. • Carlsson & Mémoli (2009), Characterization, Stability and Convergence of Hierarchical Clustering Methods. Journal of Machine Learning Research. 	<ul style="list-style-type: none"> • Otimização Convexa • Support Vector Machines • Classificação baseada em Kernel