

# EA991 - Laboratório de Aprendizado de Máquina

Métodos tradicionais de regressão

Prof. Levy Boccato  
Prof. Denis G. Fantinato



# O problema de regressão





# Relembrando

- Classificação tenta determinar a qual classe um exemplo pertence, baseado em objetos cujas classes são conhecidas, gerando um modelo que pode ser aplicado a novos objetos
- A saída esperada em classificação é um atributo nominal (classe)
- E se o valor que queremos prever for numérico, ao invés de um conjunto nominal de valores?

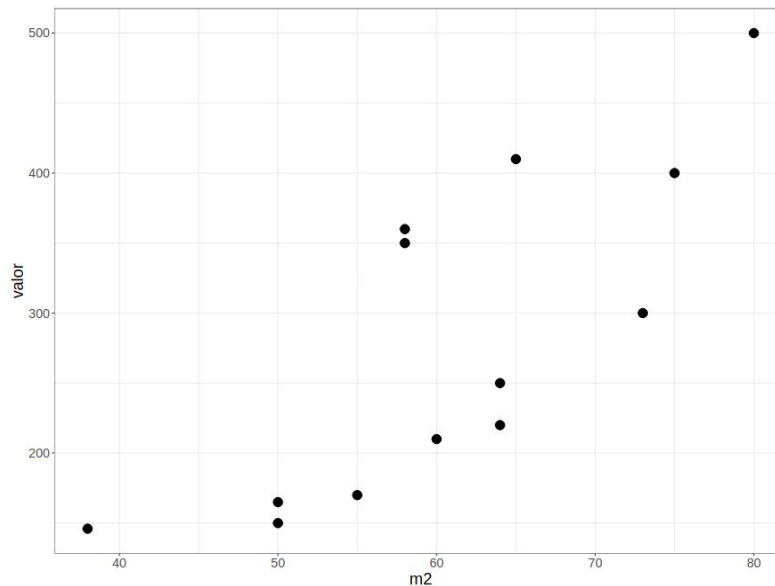
Então o nosso problema passa a ser de **regressão** ao invés de classificação!

$$\forall i, y_i \in \mathbb{R}$$

# Exemplo de Regressão

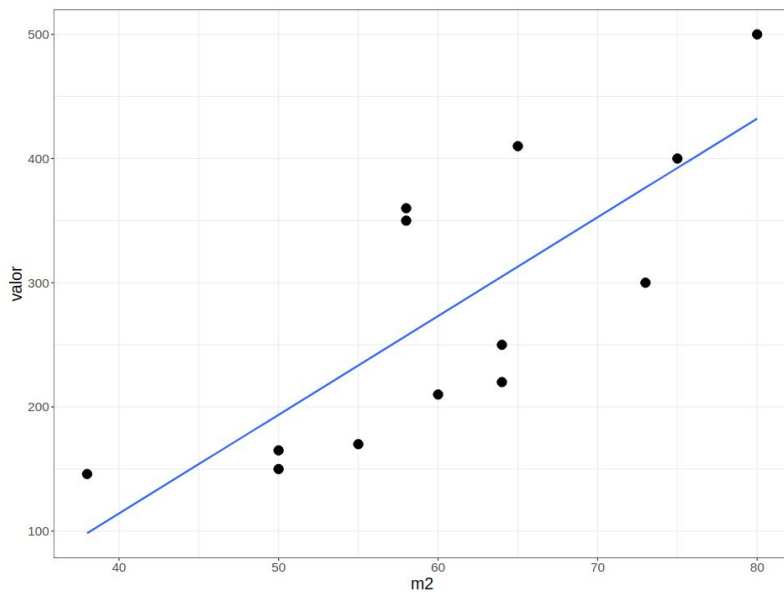
- Prever o valor (mil reais) de imóveis em função de sua área (m<sup>2</sup>):

	m2	valor
1	38.00	146.00
2	50.00	150.00
3	50.00	165.00
4	55.00	170.00
5	60.00	210.00
6	64.00	220.00
7	64.00	250.00
8	73.00	300.00
9	80.00	500.00
10	75.00	400.00
11	58.00	360.00
12	58.00	350.00
13	65.00	410.00



# Exemplo de Regressão

- Como no problema de Classificação, o modelo mais simples que podemos ter é baseado em modelos lineares em relação à entrada





# Exemplo de Regressão

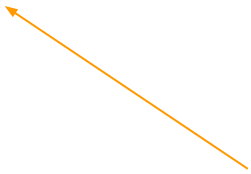
- Como no problema de Classificação, o modelo mais simples que podemos ter é baseado em modelos lineares em relação à entrada

$$g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0$$

sendo  $\mathbf{w} \in \mathbb{R}^{M \times 1}$  e  $w_0 \in \mathbb{R}$

$\mathbf{w}$  = parâmetros/pesos

$w_0$  = **bias**



Note que, diferentemente do problema de classificação, não há necessidade de uma função não linear ou um decisor.

# ***Regressão Linear***





# ***Regressão Linear***

- No problema de regressão linear, gostaríamos de obter os parâmetros do modelo de forma a minimizar uma função custo
- Opções para compor a Função Custo:

$$g(\mathbf{x}_i) - y_i$$

$$|g(\mathbf{x}_i) - y_i|$$

$$\frac{1}{2}(g(\mathbf{x}_i) - y_i)^2$$

Diferença

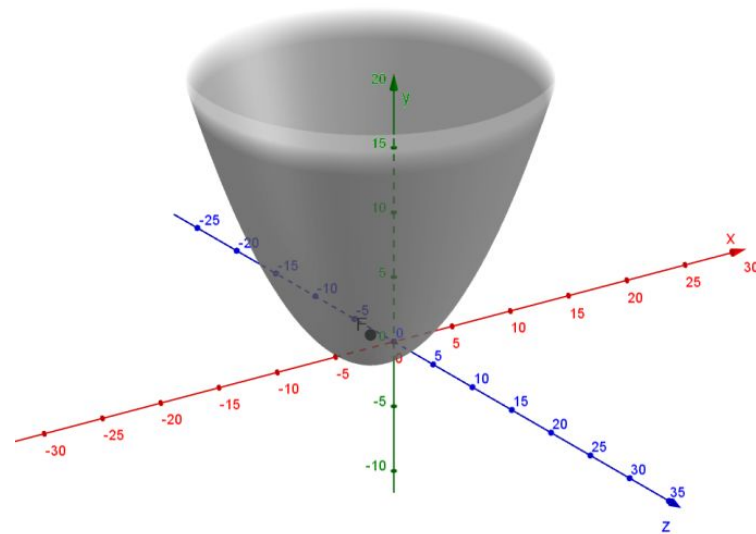
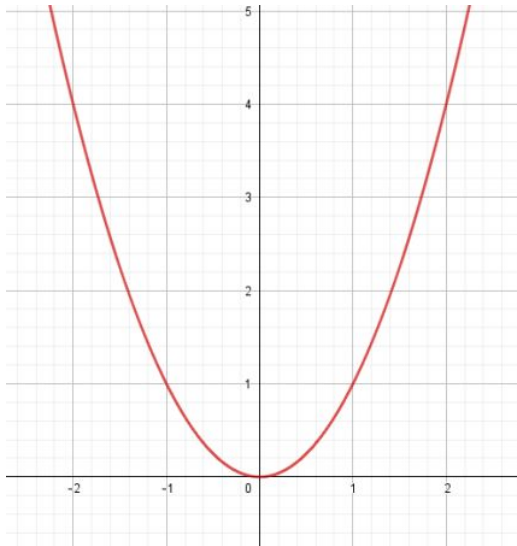
Módulo da diferença

Diferença ao quadrado



# ***Regressão Linear***

- De forma bastante interessante, a escolha da diferença ao quadrado gera uma função custo que é facilmente derivável e que possui solução fechada



# Regressão Linear

- De forma bastante interessante, a escolha da diferença ao quadrado gera uma função custo que é facilmente derivável e que possui solução fechada
- Representação do problema:

$$\phi(\mathbf{x}) = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_M \end{bmatrix}$$

Adicionamos um atributo *dummy* para corresponder ao *bias*

nesse caso  $\tilde{\mathbf{w}} = (w_0, \mathbf{w})$

e para facilitar  $\tilde{\mathbf{X}} = \begin{bmatrix} \phi(\mathbf{x}_1)^T \\ \vdots \\ \phi(\mathbf{x}_N)^T \end{bmatrix}$

Todas as amostras de treinamento são concatenadas

# Regressão Linear

- A solução para esse problema é conhecida como mínimos quadrados

$$J(\mathbf{w}) = \frac{1}{2N} \sum_{i=1}^N (y_i - g(\mathbf{x}_i^T))^2$$

$$J(\tilde{\mathbf{w}}) = \frac{1}{2N} \sum_{i=1}^N (y_i - \tilde{\mathbf{w}}^T \phi(\mathbf{x}_i^T))^2$$

$$J(\tilde{\mathbf{w}}) = \frac{1}{2N} \|\mathbf{y} - \tilde{\mathbf{X}}\tilde{\mathbf{w}}\|^2$$

$$\begin{aligned} J(\tilde{\mathbf{w}}) &= \frac{1}{2N} \|\mathbf{y} - \tilde{\mathbf{X}}\tilde{\mathbf{w}}\|^2 = \frac{1}{2N} (\mathbf{y} - \tilde{\mathbf{X}}\tilde{\mathbf{w}})^T (\mathbf{y} - \tilde{\mathbf{X}}\tilde{\mathbf{w}}) \\ &\propto \mathbf{y}^T \mathbf{y} - 2\tilde{\mathbf{w}}^T \tilde{\mathbf{X}}^T \mathbf{y} + \tilde{\mathbf{w}}^T \tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \tilde{\mathbf{w}} \end{aligned}$$

$$\nabla J(\tilde{\mathbf{w}}) = 0$$

$$2\tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \tilde{\mathbf{w}} = 2\tilde{\mathbf{X}}^T \mathbf{y}$$

$$\tilde{\mathbf{w}} = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \mathbf{y}$$

Solução fechada usou todas as entradas e saídas do conjunto de treinamento



# ***Regressão Linear***

- Entretanto, dependemos de  $X^T X$  ter inversa
  - Isso é um problema se temos atributos correlacionados
  - Ou seja, podem existir colunas de  $X$  que são linearmente dependentes
- Possíveis soluções:
  - seleção de variáveis, segundo algum critério, mantendo apenas as mais relevantes ao problema
  - redução de dimensionalidade (por exemplo, PCA)
  - usar um modelo ajustado pelo iterativamente (pelo método do gradiente)

# Modelos lineares nos parâmetros





# Modelos Lineares nos Parâmetros


- Podemos aplicar uma transformação não linear sobre os dados:

$$\phi(\mathbf{x}) = \begin{bmatrix} 1 \\ \phi(x_1) \\ \phi(x_2) \\ \vdots \\ \phi(x_M) \end{bmatrix}$$

Nesse caso, a saída do modelo será

$$g(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + w_0$$

Note que este modelo realiza uma transformação **não linear** em relação às entradas, mas é **linear** em relação aos parâmetros do modelo!





# Modelos Lineares nos Parâmetros

- Assim, se aplicarmos uma transformação não linear sobre os dados:

$$\tilde{\mathbf{X}} = \begin{bmatrix} \phi(\mathbf{x}_1)^T \\ \vdots \\ \phi(\mathbf{x}_N)^T \end{bmatrix}$$

o critério de mínimos quadrados leva à solução vista anteriormente, porém com as entradas transformadas:

$$\tilde{\mathbf{w}} = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \mathbf{y}$$

# Regularização







# Regularização

- Trata-se de uma restrição sobre os pesos do modelo, através da definição de uma distribuição *a priori*, que surge como um termo de penalidade  $P(\mathbf{w})$  na função custo:

$$\min_{\mathbf{w}} J(\mathbf{w}) + \lambda P(\mathbf{w})$$

em que  $\lambda > 0$  é chamado de coeficiente de regularização.

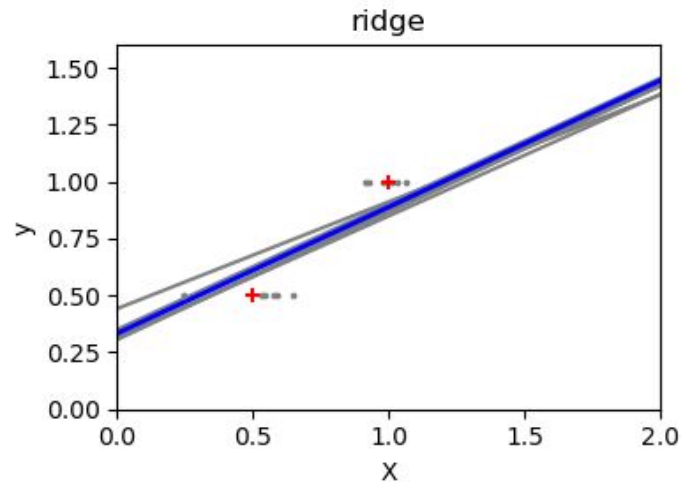
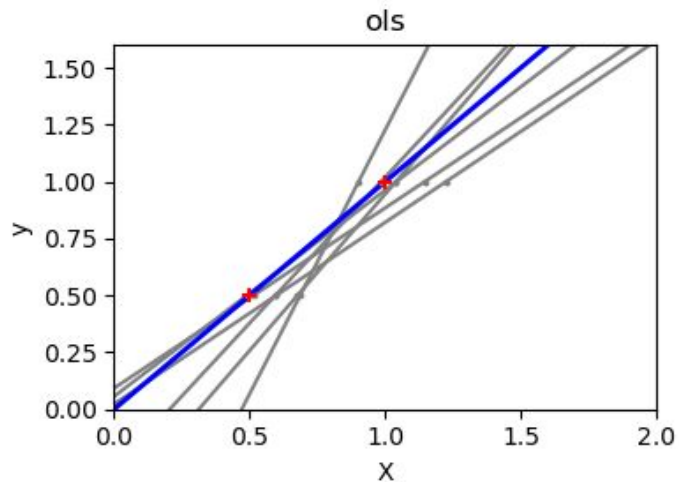
Tipos mais usados:

- Ridge -  $P(\mathbf{w})$  usa norma L2 (possui solução fechada)
- Lasso -  $P(\mathbf{w})$  usa norma L1
- ElasticNet -  $P(\mathbf{w})$  usa ambas normas L1 e L2

# Regularização

Ridge regression:

- Evita valores muito discrepantes de  $w$
- É capaz de reduzir os efeitos do sobreajuste





# Regularização

## Ridge regression:

- Evita valores muito discrepantes de  $w$
- É capaz de reduzir os efeitos do sobreajuste

## Lasso regression:

- O uso da norma L1 faz com que  $w$  tenha uma distribuição esparsa, com vários valores nulos
- Dessa forma, essa técnica pode ser utilizada para a seleção de atributos

**k-Vizinhos Mais  
Próximos**





# k-Vizinhos Mais Próximos

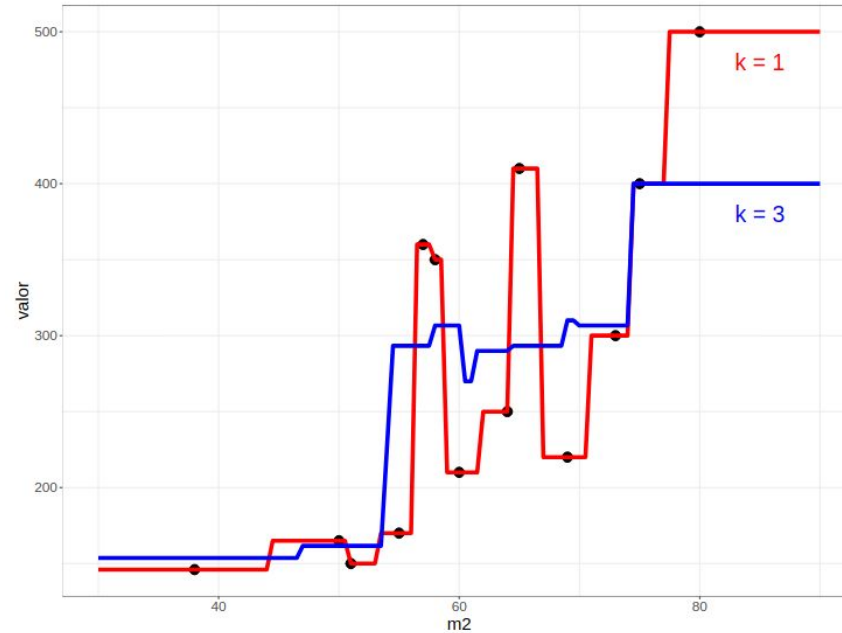
- Seja  $V$  o conjunto dos  $k$ -Vizinhos mais próximos de uma amostra de teste
- A predição para esta amostra de teste é

$$\hat{y} = \frac{1}{K} \sum_{(\mathbf{x}_i, y_i) \in V} y_i$$

- De forma análoga ao problema de classificação, essa abordagem gera um regressor que é não linear no espaço das amostras, porém é linear por partes

# k-Vizinhos Mais Próximos

- Exemplo:



# Árvores para Regressão





# Árvores para regressão

Para utilizarmos o algoritmo de Árvore para regressão, precisamos mudar duas etapas em relação ao equivalente para classificação:

- A forma de realizar a **predição numérica**
- Como **escolher o atributo** a ser utilizado para a divisão





# Árvores para regressão

Predição numérica:

- Verificamos o nó folha em que o objeto de teste se encontra
- Computamos a média dos valores de saída para cada um dos objetos no nó folha  
Seja  $F$  os objetos no nó folha

$$\hat{y} = \frac{1}{|F|} \sum_{(\mathbf{x}_i, y_i) \in F} y_i$$

Podemos ter um modelo mais sofisticado por meio de ponderação ou usando uma regressão linear nos objetos da folha



# Árvores para regressão

## Escolha do Atributo:

- Como a predição é feita considerando a média, selecionamos a divisão que produz maior redução no erro quadrático (ao invés do índice de gini ou entropia)
- Seja  $S$  um (sub)conjunto dos dados, podemos indicar o erro quadrático médio obtido pela predição da média como:

$$\text{EQM}(\mathcal{S}) = \frac{1}{N} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}} (y_i - \bar{y})^2 \quad \bar{y} = \frac{1}{N} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}} y_i$$

$$\text{EQM\_DIV}(D(\mathcal{X}, A)) = \sum_{\mathcal{X}_i \in D(\mathcal{X}, A)} \frac{|\mathcal{X}_i|}{|\mathcal{X}|} \text{EQM}(\mathcal{X}_i)$$

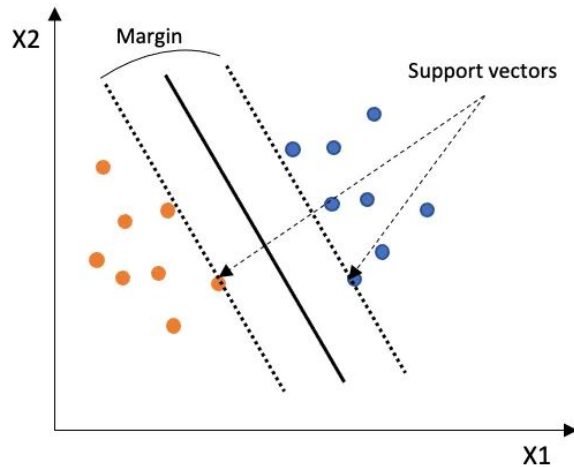
$$\text{RE}(\mathcal{X}, A) = \text{EQM}(\mathcal{X}) - \text{EQM\_DIV}(D(\mathcal{X}, A))$$

# Máquinas de Vetores Suporte para Regressão

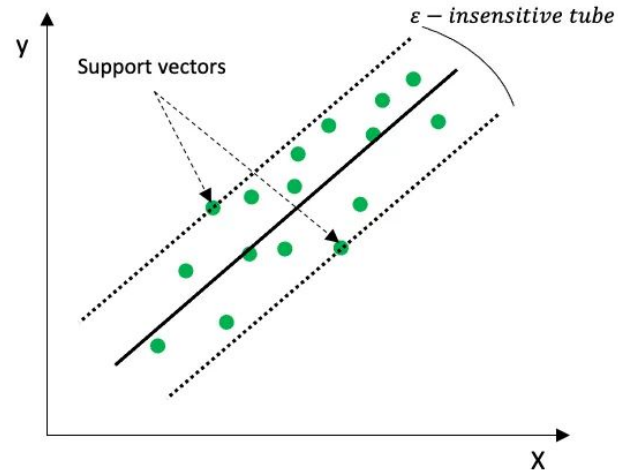


# Support Vector Regression (SVR)

- No SVR, a margem é definida como a tolerância ao erro do modelo, também chamada de tubo  $\epsilon$ -insensível.



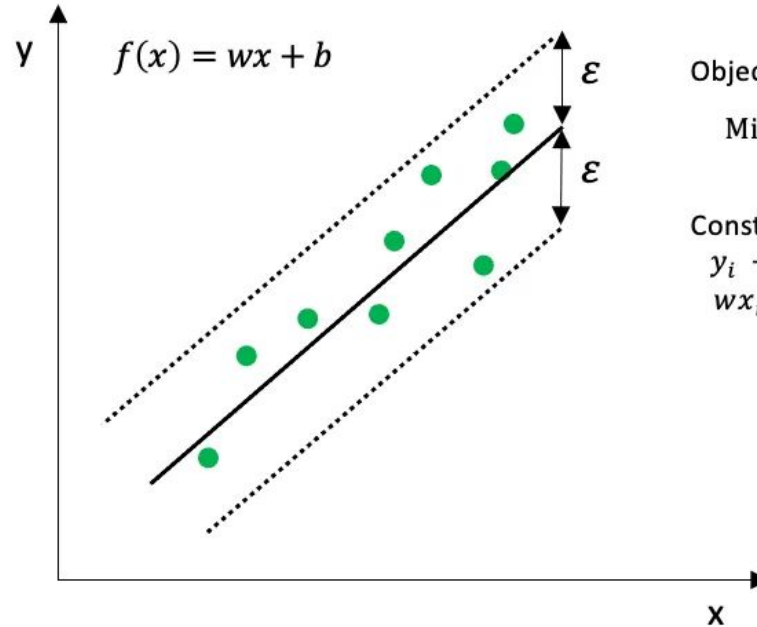
Classification problem using SVM



Regression problem using SVR

# Support Vector Regression (SVR)

- Este tubo permite algum desvio dos pontos de dados em relação ao hiperplano sem ser contabilizado como erro.
- A formulação é similar ao problema de classificação, porém as amostras devem estar dentro do tubo  $\varepsilon$ -insensível



Objective:

$$\text{Minimize: } \frac{1}{2} \|w\|^2$$

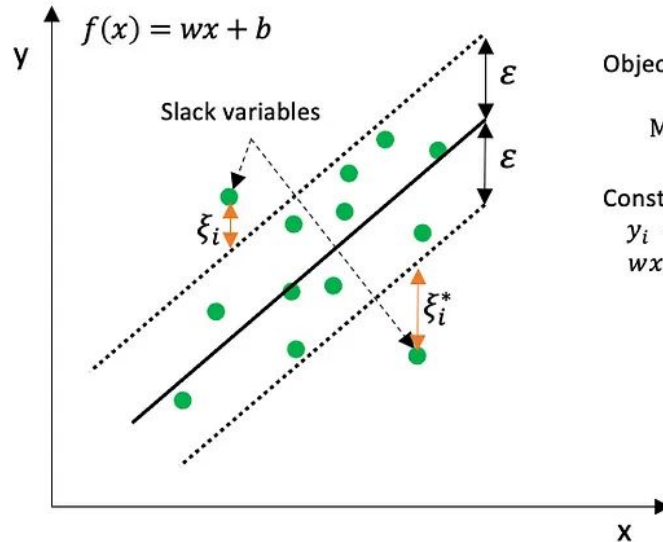
Constraints:

$$y_i - wx_i - b \leq \varepsilon$$

$$wx_i + b - y_i \leq \varepsilon$$

# Support Vector Regression (SVR)

- Analogamente ao problema de classificação, também é possível admitir uma tolerância a erros (amostras fora do tubo  $\varepsilon$ -insensível)
- Além disso, permite que se obtenha uma regressão não linear.



Objective:

$$\text{Minimize: } \frac{1}{2} \|w\|^2 + c \sum_{i=1}^l (\xi_i + \xi_i^*)$$

Constraints:

$$\begin{aligned} y_i - wx_i - b &\leq \varepsilon + \xi_i \\ wx_i + b - y_i &\leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* &\geq 0 \end{aligned}$$

# Mensurando Acertos/Erros





# Mensurando Acertos/Erros

- No contexto de regressão faz sentido vermos a diferença entre o valor predito e o valor real. Nesse âmbito, as duas métricas mais usuais são:
- Erro quadrático médio (Mean Squared Error):

$$\text{MSE}(\mathcal{S}, g) = \frac{1}{N} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}} (y_i - g(\mathbf{x}_i))^2$$

- Root Mean Square Error

$$\text{RMSE}(\mathcal{S}, g) = \sqrt{\frac{1}{N} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{S}} (y_i - g(\mathbf{x}_i))^2}$$





# Mensurando Acertos/Erros

- Coeficiente de determinação  $R^2$ :

$$R^2 = 1 - \frac{\sum (y_{\text{actual}} - y_{\text{predicted}})^2}{\sum (y_{\text{actual}} - \bar{y})^2}$$

$\bar{y}$  é a média dos valores  $y_{\text{actual}}$ .

- $R^2 = 1$ : o modelo explica perfeitamente toda a variância na variável-alvo.
- $R^2 = 0$ : o modelo não explica nenhuma variância; as previsões não são melhores do que simplesmente usar a média.
- $R^2 < 0$ : o modelo tem um desempenho pior do que simplesmente usar a média, indicando um ajuste ruim

Existem outras medidas, mas normalmente são variações dessas.