Análise Preditiva
Aula 4

Bruno Silva
sbruno@br.ibm.com

Bibliografia

- Hands-On Data Science and Python Machine Learning. Frank Kane. Packt Publishing. 2017
- Advanced Data Analytics Using Python: With Machine Learning, Deep Learning and NLP Examples. Sayan Mukhopadhyay. Apress. 2008
- Python: Advanced Predictive Analytics. oseph Babcock; Ashish Kumar. Packt Publishing, 2017

Regressão Logística

- Regressão linear assume que existe um relacionamento linear entre entrada e saída
- Na aula de hoje veremos:
  - Conceitos matemáticos relacionados a regressão logística
  - Implementar regressão logística com *python*
  - Observar métricas para validação de modelos

Regressão Linear X Regressão Logística

- Regressão linear
  - Previsão números
- E se quiséssemos prever categorias ao invés de números?
- Regressão logística
  - Saída da previsão é uma categoria (geralmente binária)

Exemplos Regressão Logística

- Prever se um cliente vai comprar um carro ou não?
- Prever se um time vai ganhar ou perder um jogo?

Comparação Regressão Linear X Logística

	Linear regression	Logistic regression
Predictor variables	Continuous numeric/categorical	Continuous numeric/categorical
Output variables	Continuous numeric	Categorical
Relationship	Linear	Linear (with some transformations)

Entendendo a Matemática da LR

- Conceito de probabilidade condicional
- Conceito de chance de sucesso (odds)
- Diferença *Logistic Regression* e *Linear Regression*

Probabilidade Condicional

Defining Conditional Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Probabilidade Condicional

Let:

- M -> The person is male
- F -> The person is female
- B -> The person bought a product
- D -> The peron did not buy a product

Then the conditional probabilities are:

- Probability of buying males:

$$P(B|M) = \frac{P(B \cap M)}{P(M)}$$

- Probability of not buying males:

$$P(B|M) = \frac{P(D \cap M)}{P(M)}$$

- Probability of buying females:

$$P(B|M) = \frac{P(B \cap F)}{P(F)}$$

- Probability of not buying females:

$$P(B|M) = \frac{P(D \cap F)}{P(F)}$$

Católica de Santa Catarina
Centro Universitário

Odds

Odds

Define a success rate for a desired event

- Odds of purchase by males =
$$\frac{P(B|M)}{P(D|M)} = \frac{P(B|M)}{1 - P(B|M)} = \frac{0.49}{0.51} = 0.96$$
- Odds of purchase by females =
$$\frac{P(B|F)}{P(D|F)} = \frac{P(B|F)}{1 - P(B|F)} == \frac{0.60}{0.40} = 1.5$$

Odds Between Groups

Odds between groups ¶

One better way to determine which group has better odds of success is by calculating odds ratios for each group. The odds ratio is defined as follows:

$$OddsRation = \frac{OddsGroup1}{OddsGroup2}$$

Then:

$$OddsRatio(males) = \frac{OddsMales}{OddsFemales} = \frac{0.96}{1.5} = 0.64$$

$$OddsRatio(females) = \frac{OddsFeales}{OddsMales} = \frac{1.5}{0.96} = 1.54$$

Linear Regression analogy

Remember linear regression equation:

$$Y = \beta_0 + \beta_1 * X + \epsilon$$

- X can assume any value in range $-\infty, +\infty$. Therefore, it is hard to properly match these values in a $[0, 1]$ range
- What if we try to predict the probabilities associated with the two events rather than the binary outcomes? Predicting the probabilities will be feasible as their range spans from 0 to 1.

$$P(Y) = a + b * X$$

- The range problem persists, P $[0, 1]$ while X $[-\infty, +\infty]$

What if we use the odds instead of P, the range would be $[0, \infty]$

$$P/(1 - P) = a + b * X$$

What if we use the log of odds?

$$log(P/(1 - P)) = a + b * X$$

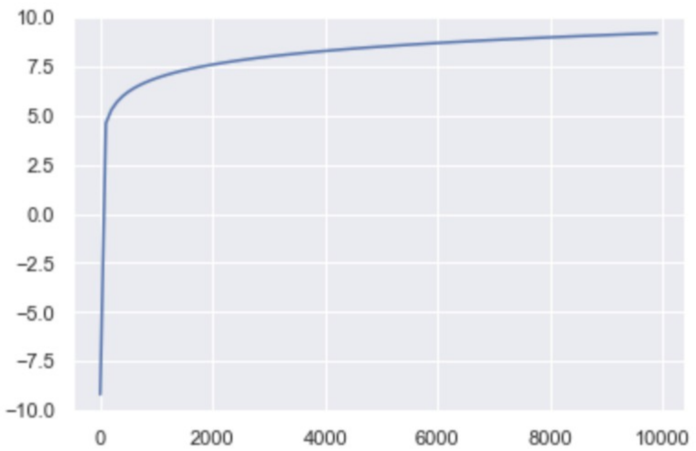
Imagem função logarítmica

In [65]:
```
# Observe logarithm can assume any value in -infinite to infinite
from matplotlib import pyplot as plt
import seaborn as sns

sns.set()
X = np.arange(10**-4, 10**4, 100)
Y = np.log(X)


plt.plot(X, Y)
```

Out[65]:
```
[<matplotlib.lines.Line2D at 0x161e0fa90>]
```



Equações

$$log(P/(1-P)) = a + b * X$$

$$\frac{P}{1-P} = e^{a+b*X}$$

$$P = (1-P) * e^{a+b*X}$$

$$P = e^{a+b*X} - P * e^{a+b*X}$$

$$P + P * e^{a+b*X} = e^{a+b*X}$$

$$P(1 + e^{a+b*X}) = e^{a+b*X}$$

$$P = \frac{e^{a+b*X}}{1 + e^{a+b*X}}$$

$$P = \frac{1}{1 + e^{-(a+b*X)}}$$

Transformações Aplicadas

Transformation (LHS)	Range of LHS	Range of LHS
Y	*Y= 0* or *Y= 1*	*Infinity<X<+infin-ity*
P (Probability)	*0<P<1*	*Infinity<X<+infin-ity*
P/1-P (Odds)	*0<P/1-P<+infinity*	*Infinity<X<+infin-ity*
log(P/1-P)	*-infinity<log(P/1-P)<+in-finity*	*Infinity<X<+infin-ity*

Caso com múltiplas *Features*

For a multiple logistic regression, the equation can be written as follows:

$$\log(P/1-P) = a + b_1 * X_1 + b_2 * X_2 + b_3 * X_3 + \dots\dots\dots\dots\dots\dots + b_n * X_n$$

$$P = \frac{1}{1+e^{-(a+b_1*X_1+b_2*X_2+b_3*X_3+\dots+bn*Xn)}}$$

If we replace (*X1*, *X2*, *X3*,…,*Xn*) with *Xi'* and (*b1*, *b2*, *b3*,----,*bn*) with *bi'*, the equation can be rewritten as follows:

$$P = \frac{e^{a+b*Xi'}}{1+e^{a+b*Xi'}} = \frac{1}{1+e^{-(a+b*Xi')}}$$

Plot da regressão logística

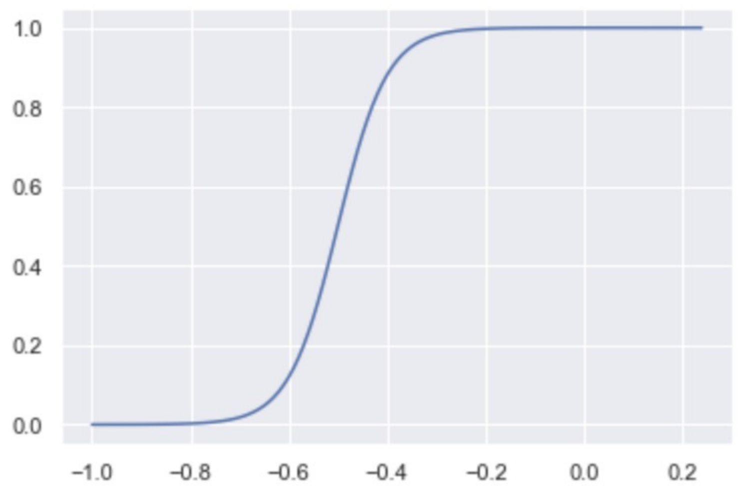
In [77]:
```
# Lets see it in a graph

a, b = 10, 20
logistics = lambda x: 1 / (1 + np.e**-(a + b*x))
vlogistcs = np.vectorize(logistics)

X = np.arange(-1, 0.25, 0.01)
Y = vlogistcs(X)

plt.plot(X, Y)
```

Out[77]: `[<matplotlib.lines.Line2D at 0x163e49a58>]`



Quick Question

3/3 points (graded)

Suppose the coefficients of a logistic regression model with two independent variables are as follows:

$$\beta_0 = -1.5, \quad \beta_1 = 3, \quad \beta_2 = -0.5$$

And we have an observation with the following values for the independent variables:

$$x_1 = 1, \quad x_2 = 5$$

What is the value of the Logit for this observation? Recall that the Logit is log(Odds).

What is the value of the Odds for this observation? Note that you can compute e^x, for some number x, in your R console by typing exp(x). The function exp() computes the exponential of its argument.

What is the value of P(y = 1) for this observation?

Exemplo de Regressão Logística

- HealthCare Example

Métricas de Avaliação de Erro

- A saída de uma regressão logística é uma probabilidade
- Então as decisões são baseadas na probabilidade de saída
- *Ex. P(PoorCare = 1) >= t*
- t representa um limiar
- Qual valor de t utilizar?

Métrica de Avaliação de Erros

- Deve-se utilizar um valor de t onde os erros são “melhores”
- Se *t* é grande *PoorCare = 1* raramente vai ocorrer
  - Mais falsos negativos
- Se *t* é pequeno *PoorCare = 0* raramente vai ocorrer
  - Mais falsos positivos
- Quando não se prefere um grupo em relação a outro, escolhe-se t = 0.5

Matriz de confusão

	Predicted = 0	Predicted = 1
Actual = 0	True Negatives (TN)	False Positives (FP)
Actual = 1	False Negatives (FN)	True Positives (TP)

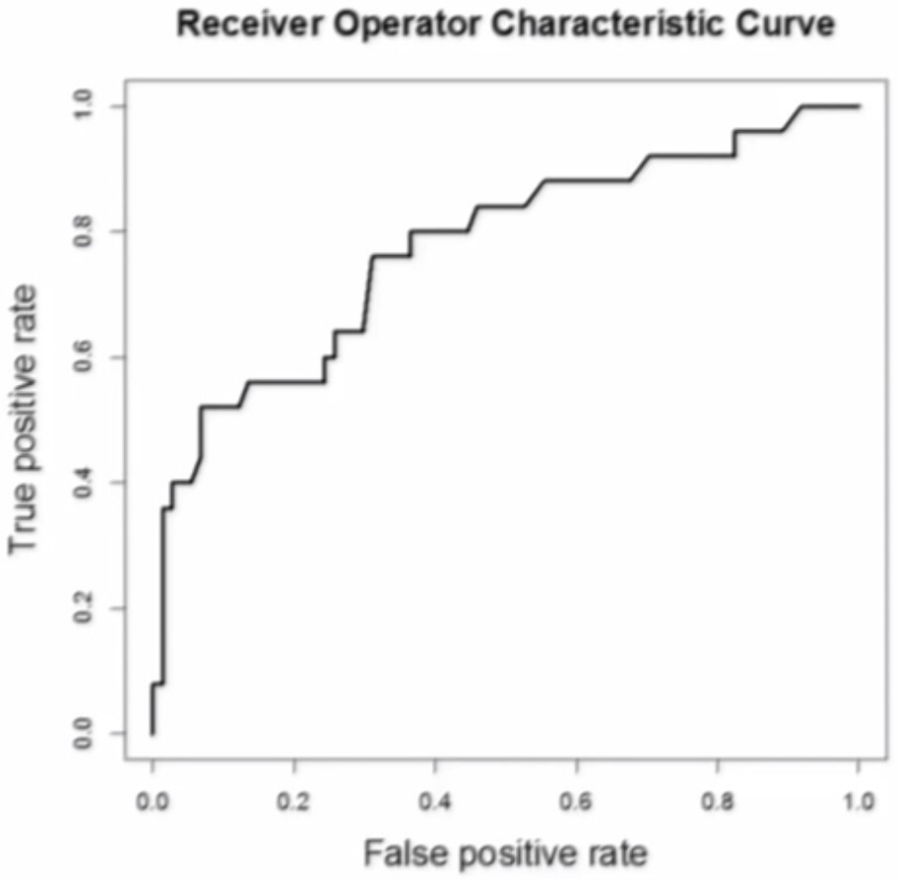
Métricas

- Sensitivity (True Positive Rate) = TP / (TP + FN)
- Specificity (True Negative Rate) = TN / (TN + FP)

O que ocorre com modelos com altos valores de t? e baixos?

Receiver Operator Characteristics (ROC) Curve

- True positive rate (sensitivity) on y-axis
  - Proportion of poor care caught
- False positive rate (1-specificity) on x-axis
  - Proportion of good care labeled as poor care



ROC Curve