Modelo de Classificação
Gradient Boosting

Mercado Imobiliário

Alan Leal – CEDEPLAR/UFMG e FEA/USP

30/11/2022

Entendendo a necessidade de um modelo de classificação

Os imóveis são negociados em mercado próprio ou mercado de terrenos?

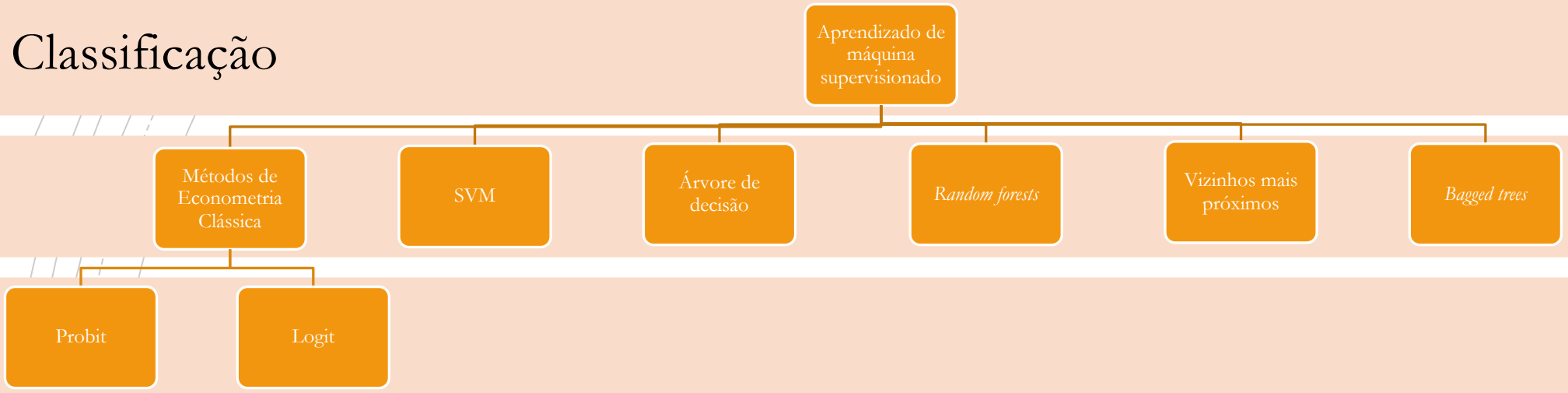
O que em termos gerais poderia caracterizar o mercado de lotes e o mercado imobiliário próprio de cada imóvel?

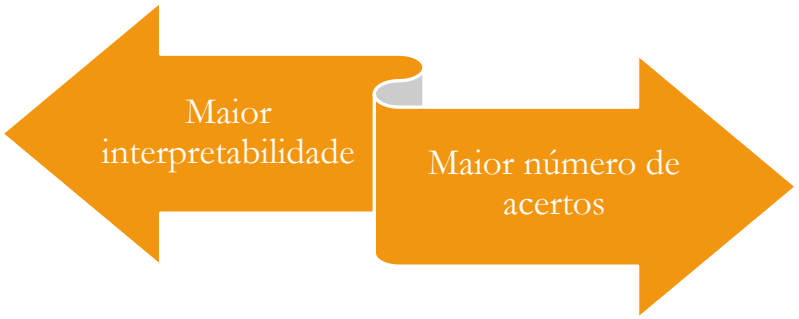
Como identificar na prática e em massa os imóveis que são negociados em mercado próprio ou no mercado de terrenos?

Principais métodos e possibilidades

Métodos de classificação

Tarefa de Classificação





Trade-off entendimento vs acertos de previsão

Trade-off entre interpretabilidade e previsibilidade

Métodos de classificação caixa-preta

Implementação do LOGIT: breves comentários

Implementação do método de *bagged trees* via algoritmo de *Gradient Boosting*

LOGIT

Modelo de previsão de probabilidade condicional.

Seja Y* uma variável latente, Y={0,1} e X um vetor de covariadas.

Defina $Y = I\{Y^* \geq 0\}$. E defina adicionalmente uma relação entre Y e X da seguinte forma:

$Y^* = X\beta + u$

Então, $Y = P(Y^* \geq 0|X) = P(X\beta + u \geq 0) = P(u \geq -X\beta|X) = P(u \leq X\beta) = G(\beta X)$

Se u é simétrica

No caso do modelo PROBIT, assume-se que G é uma função normal padrão.

No caso do modelo LOGIT, assume-se que G é uma função sigmoide logística.

LOGIT

- O LOGIT, assim como o PROBIT, pode ser resolvido através do método de máxima verossimilhança
- Seja Y uma variável binária de valores observados que assume valores {0,1} apenas. Seja X uma matrix nxk, na qual a primeira coluna é composto de 1's.
- A verossimilhança de cada observação Y_i é dada por:

$$L(\beta_i, y_i, x_i) = [\Phi(x_i\beta)]^{y_i}[1 - \Phi_{x_i\beta}]^{1-y_i}$$

- A log-verossimilhança, por sua vez, será dada por:

$$l(\beta, y, X) = \sum_{i=1}^{N}[-ln(1 + exp(x_i\beta)) + y_i x_i \beta] \quad (1)$$

- O estimador de máxima verossimilhança é tal que a equação (1) é maximizada.

SVM

O *Suport Vector Machine* também é um método de classificação bastante engenhoso que tende a produzir resultados melhores de classificação que o LOGIT, contudo ele me média acerta menos que o método de *bagged trees*.

O método intuitivamente constrói um hiperplano separador das observações num espaço multidimensional. O espaço tem K+1 dimensões, em que K é o número de variáveis preditoras e a dimensão extra dá conta justamente da variável a ser classificada.

No exercício aqui empreendido, isso consiste na variável binária que diz respeito ao fato de o imóvel estar sendo negociado no mercado de terrenos ou não.

SVM

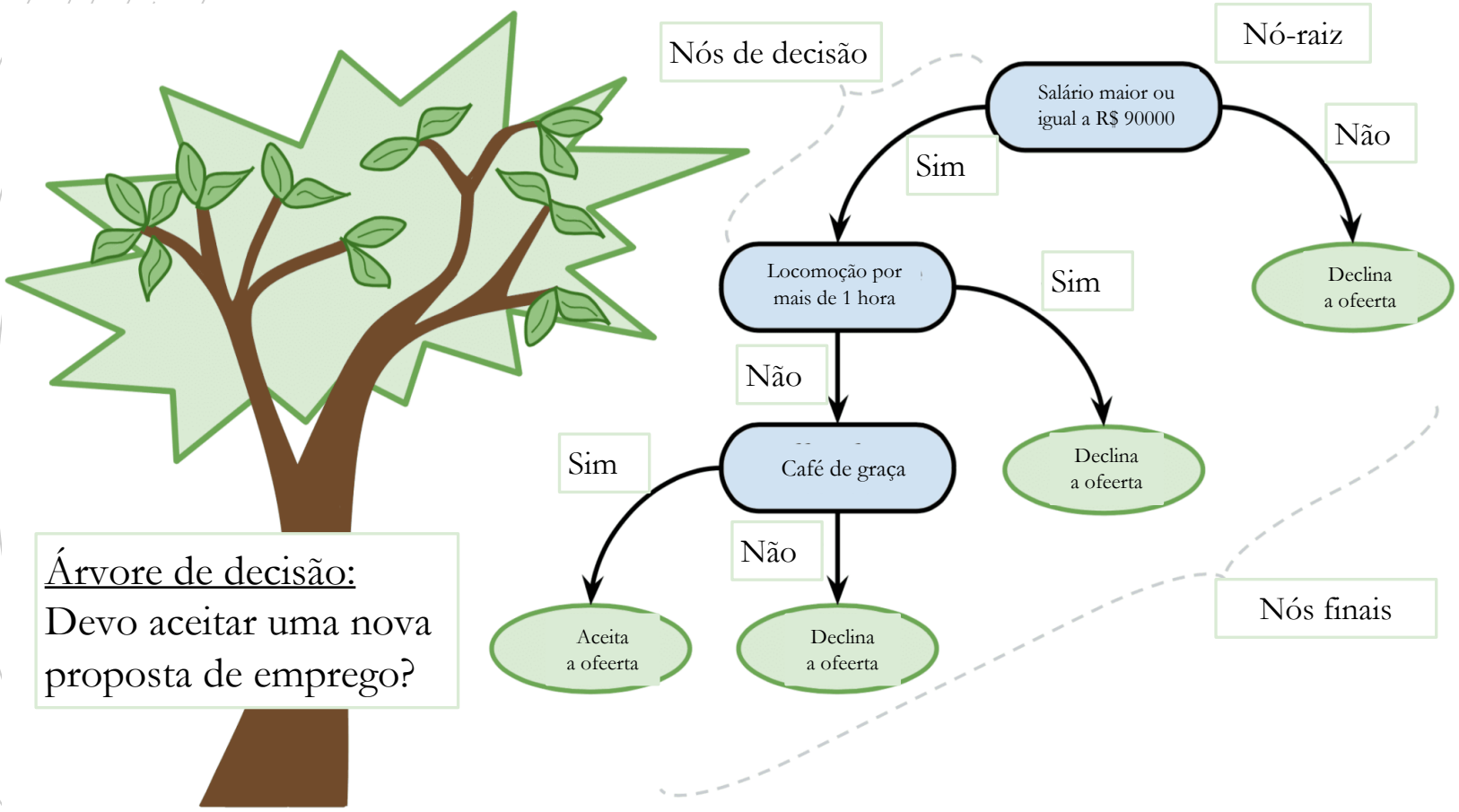
O *Suport Vector Machine* também é capaz de classificar dados que não são separáveis, isto é, ele também consegue lidar com observações que não são facilmente diferenciadas por um hiperplano separador. Para contextualizar com profundidade, esse processo ocorre através de um processo de maximização.

Aqui derivamos a matemática por trás desse classificador no caso de dados separáveis, que por simplicidade é o mais facilmente compreensível.

Gradient Boosting

O Gradient Boosting é um algoritmo veloz e rápido que trabalha com *bagged trees*.

Uma árvore de decisão simples tem o seguinte formato:



Gradient Boosting

Árvores de decisão mais complexas não têm visualizações fáceis quanto a presente no último slide.

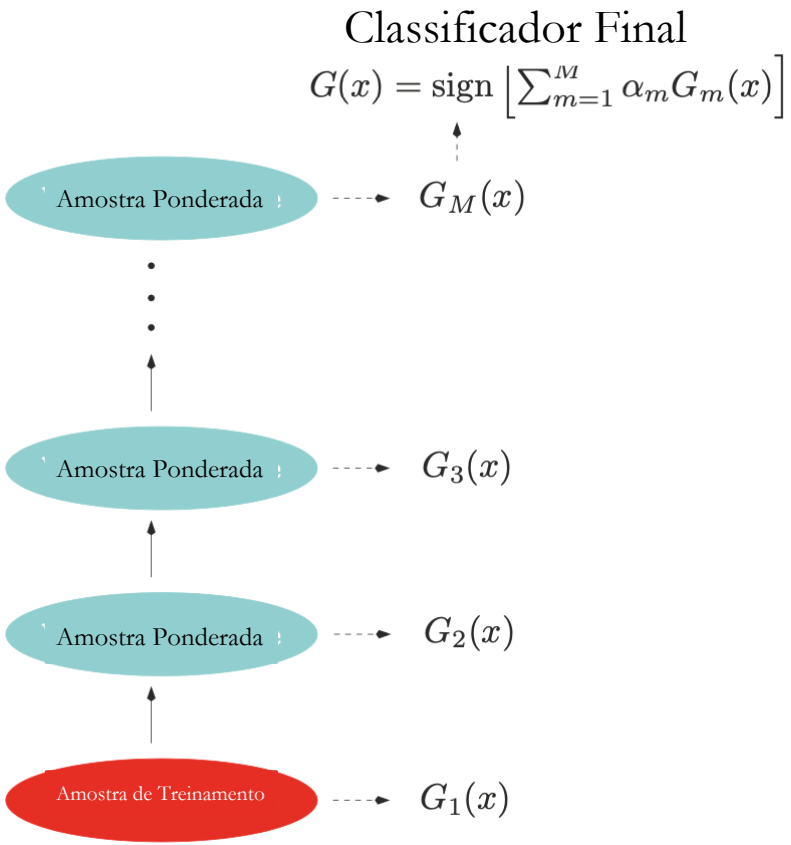
Contudo, a regra de decisao de uma árvore pode ser usada para qualquer nova observação.

No exemplo anterior, se dois indivíduos têm a mesma preferência por salário, transporte e café de graça, uma regra de decisão de um deles pode ser utilizada para outro.

O Gradient Boosting também constrói uma árvore de decisão com base na amostra de treinamento. Assim, ele constrói regras de decisão sucessivas cujo objetivo é retornar no nosso caso um valor de 1 ou 0, indicando se o imóvel está sendo negociado no mercado de lotes ou em seu mercado próprio. Ele pode ser usado também em tarefas de regressão, isto é, tarefas cuja variável *target* é contínua.

Adicionalmente, o treinamento da base de casas, galpões e barracões ajuda a visualizar se um novo imóvel por suas características têm chances de estar se associando a um projeto no futuro e dessa forma ter sido negociado na verdade no mercado de lotes.

ADABoost: Desenho Esquemático



Fonte: The elements of statistical Learning, Hastie et al. (2009)

Algoritmo *Gradient Boosting* para classificação dos imóveis

- Intuição por trás do Gradient Boosting
- Implementação do algoritmo ADABoost:

1. Inicie a classificação com pesos idênticos para cada subamostra e replique os seguintes passos por N vezes (no nosso caso 100):
    1. Classifique os dados com base nos pesos atribuídos
    2. Meça o erro de classificação
    3. Calcule o log-erro da sua tarefa de classificação atual
    4. Use o log-erro (ou outra métrica de erro) para atualizar seus pesos, de modo a dar maior peso na tarefa de classificação às amostras erroneamente classificadas
    5. Pule para a n+1 classificação
2. Retorne a árvore de classificação

Implementando na prática *Gradient Boosting*

- Variável dependente: intuição e construção

1. Junção do Índice Cadastral da base do ITBI à base de projetos. Por que realizar este exercício?
2. Espera-se que imóveis associados a algum projeto tenham maior probabilidade de estarem sendo negociados no mercado imobiliário de terrenos e não exatamente no mercado próprio de sua tipologia.
3. Dessa forma, a variável de classificação no modelo (variável match) consiste numa variável binária assumindo o valor 1, caso o imóvel do ITBI esteja associado a algum projeto, e 0, caso contrário.
4. Previsão e uso futuro do modelo para novos dados imobiliários do ITBI.

Variáveis utilizadas no exercício empírico

- Variáveis:

1. Valor declarado do imóvel
2. Ano de transação
3. Padrão de acabamento (sob a forma de cinco dummies)
4. Área construída adquirida
5. Área do terreno adquirida
6. Idade do imóvel
7. IQVU
8. Distância do imóvel ao hospital mais próximo
9. Distância do imóvel a praças regionais
10. Média dos imóveis vizinhos que são apartamentos (wAP)
11. Média da idade dos imóveis vizinhos (widade)
12. Interação entre wAP e widade
13. Dummy_app: dummy se o imóvel está ou não em área de proteção ambiental

Implementação Prática

- Divisão da Amostra
- Implementação em várias linguagens: R, Python, MATLAB, Julia, dentre outras.

Métricas de análise do ajuste do modelo

Métrica	Acurácia	AUC	*Matthews correlation coefficient*
Fórmula	$\frac{C}{N}$	Área sobre a curva ROC	$\frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
Significado	C é o número de casos corretos, enquanto N é o número total de casos classificados. A acurácia mede assim a média de acertos do modelo, sem distinção entre casos positivos e negativos	A curva ROC (*receiver operating characteristic curve*) plota a sensistividade contra a taxa de falsos positivos ($\frac{FP}{FP+TN}$). Quanto maior a área sob a curva ROC, maior a precisão do modelo. Em termos práticos, quanto maior o valor de AUC (*area under curve*), maior a chance de que um valor positivo aleatório tenha maior probabilidade predita do que um valor negativo aleatório	TP, TN, FP e FN são, respectivamente, os verdadeiros positivos, verdadeiros negativos, os falsos positivos e falsos negativos, obtidos via matriz de confusão. *Matthews correlation coefficient* mede o acerto do modelo de classificação levando em conta o possível desbalanceamento de classes presentes no modelo. Desse modo, ele é mais confiável que a acurácia

Exercício Empírico

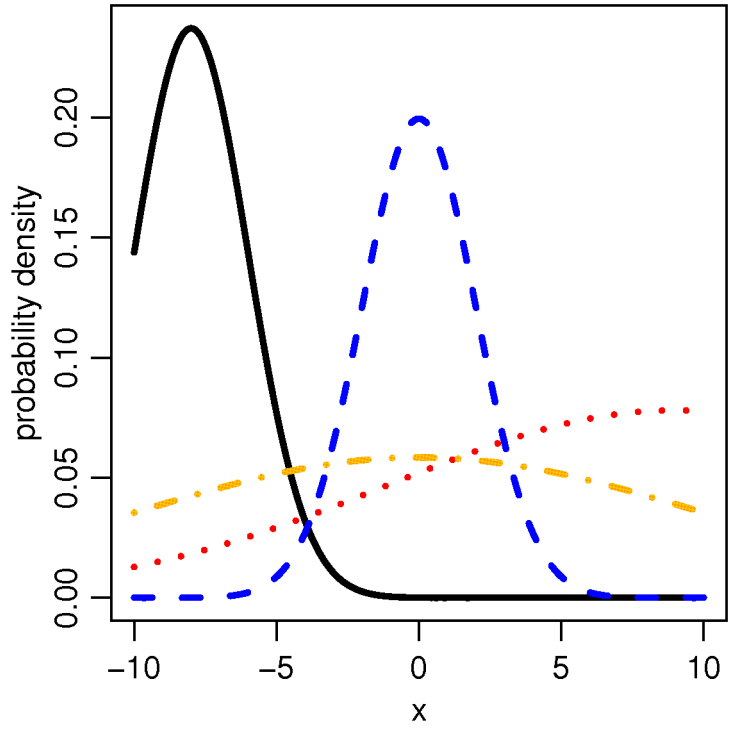
Como forma de explicitar o conteúdo visto aqui, vamos rodar o modelo de classificação para um subconjunto da base de dados de casas no Pyhton.
Também vamos considerar o modelo LOGIT, SVM e Gradient Boosting nas nossas métricas de acurácia e acerto.
Além dessa apresentação, o jupyter notebook contendo o código do modelo de classificação se encontra disponível, sob solicitação, através do e-mail: alanleal@usp.br

Exercício Empírico

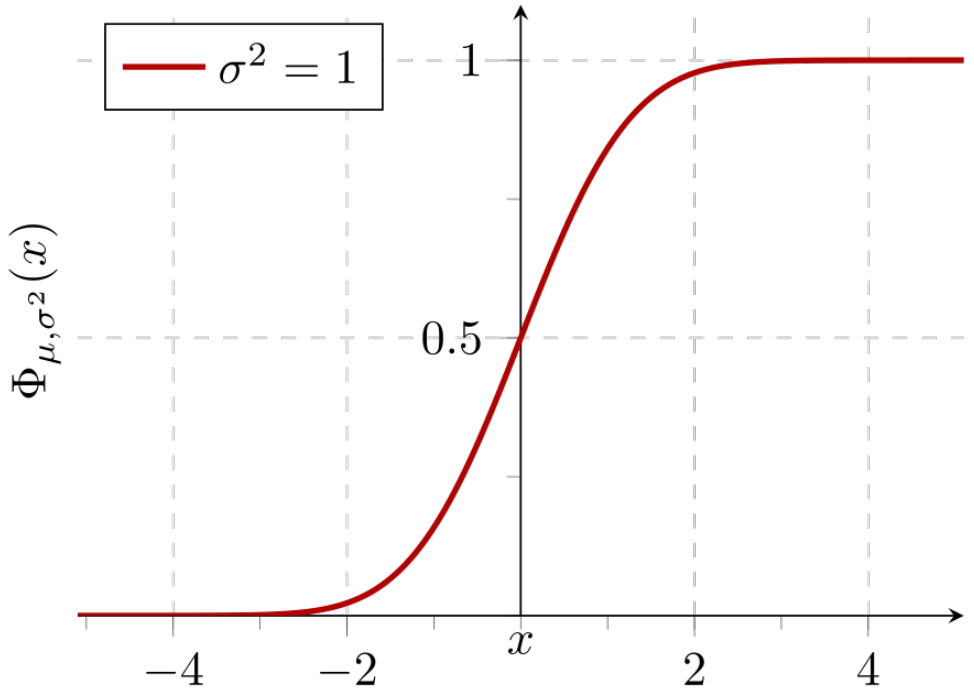
As métricas de acerto no caso do Gradient Boosting todas superaram os 90% na amostra de teste, mas o que acontece com as observações classificadas erroneamente ou como somos capazes de antecipar um erro de classificação enquanto ele acontece?
Algumas estatégias foram empregadas para corrigir possíveis erros, tal como olhar para imóveis precificados com valores relativos muito baixos ou imóveis precificados com valores muito altos.
Essa correção *ad hoc* garante de certa forma que os erros de classificação não contaminassem o modelo de preços. Apesar de o erro ser de apenas um dígito, essas medidas de certa forma conferiram maior robustez ao modelo de classificação e ao modelo de precificação de imóveis desenvolvido pelo CEDEPLAR/UFMG em parceria com a Prefeitura de Belo Horizonte.

Distribuição Normal Padrão

Função de densidade



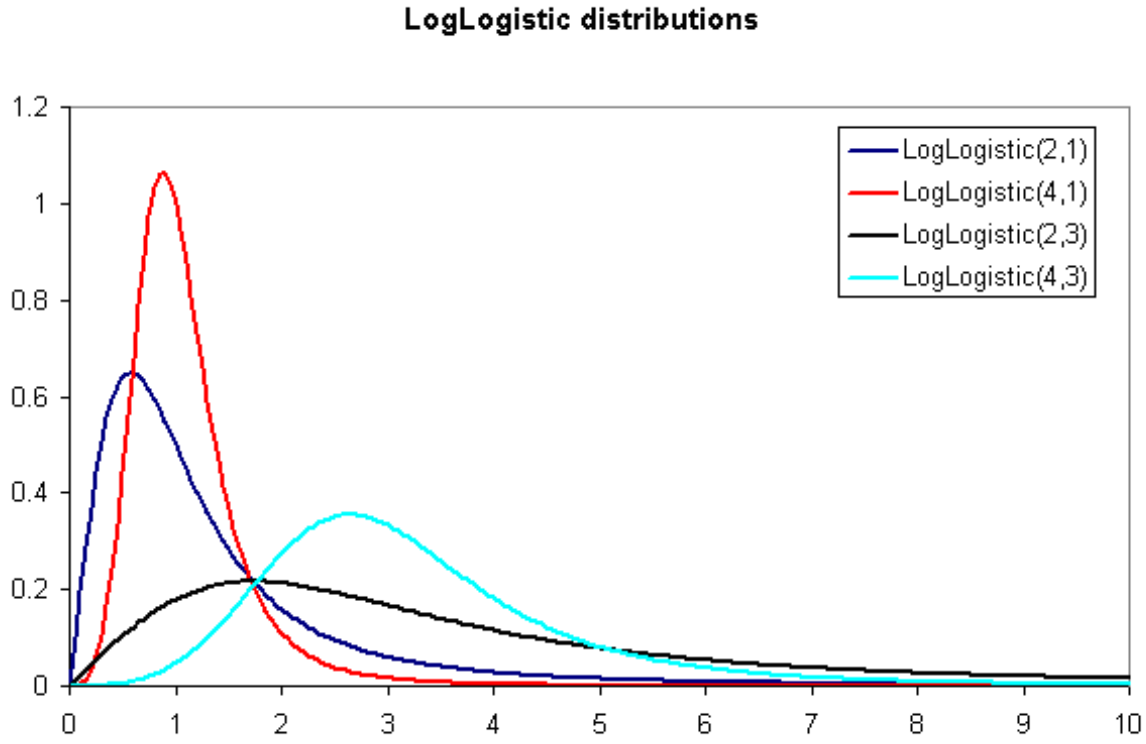
Função de distribuição (acumulada)



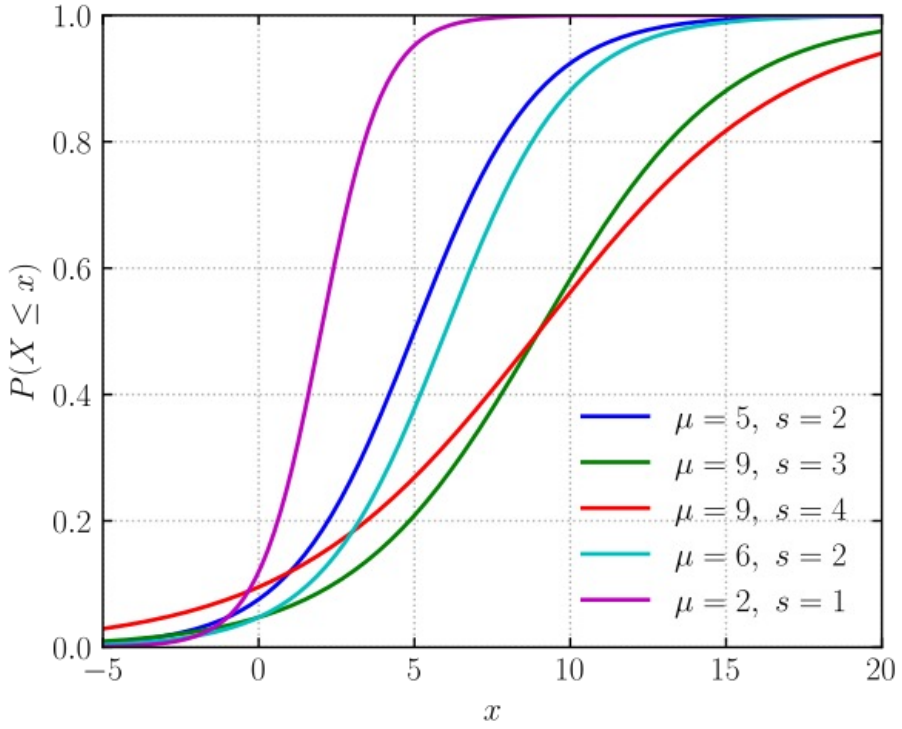
Voltar

Distribuição Sigmoide Logística

Função de densidade



Função de distribuição (acumulada)



Voltar

Máxima Verossimilhança

- O estimador de máxima verossimilhança é um estimador baseado em um processo de otimização (esses estimadores são chamados de M-estimadores). A função log-verossimilhança nada mais é do que o logaritmo da verossimilhança. A função a ser otimizada no caso do log verossimilhança é o somatório das log-verossimilhanças amostrais (ou o produto das verossimilhanças, o que é mais complexo analítica e computacionalmente).
- Na Econometria, contudo, geralmente não estamos interessados em estimar parâmetros de distribuições incondicionais, mas sim condicionais. Dessa forma, postula-se uma relação entre a variável dependente e as explicativas em termos de parâmetros numa distribuição condicional.
- O método de estimação de parâmetros de interesse pela máxima verossimilhança condicional é consistente e normalmente assintótico. Na verdade, ele é o método demonstradamente que produz estimativas para a variância que alcançam o limite inferior de Cramer-Rao. Dessa forma, todos os outros estimadores quando possíveis são comparados a ele. Um ponto negativo desse estimador é que realiza hipóteses bastante fortes sobre a distribuição adjacente ao processo gerador de dados, sendo, pois, um método puramente paramétrico, o que nem sempre é interessante.

Voltar

Probabilidade de uma variável aleatória dicotômica

Seja X uma variável que assume valores {0,1}, tal qual uma moeda que assume valores Cara e Coroa. Então, a densidade dessa variável aleatória dicotômica será dada por:

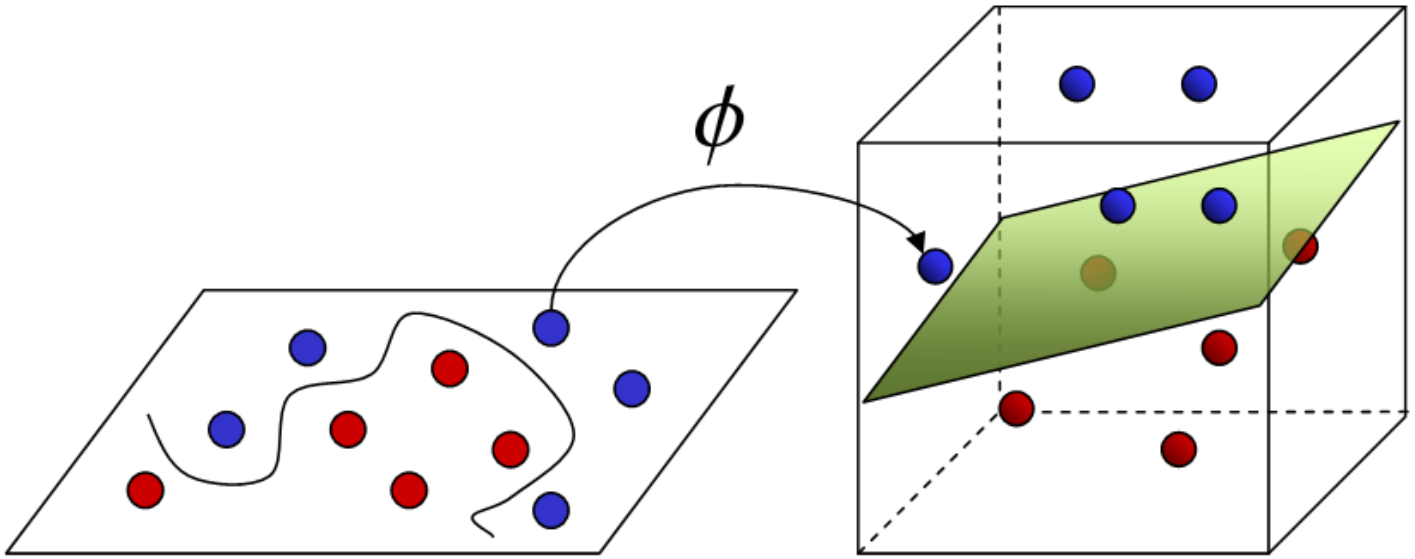
$$f(k;p) = p^k(1-p)^{1-k}, \text{ for } k \in \{0,1\}$$

No caso do LOGIT, temos que: $p = \Phi(x_i\beta)$ e $y_i = k$

Voltar

Visualização do SVM-I

A figura adiante ajuda a compreender a ideia de separar as observações por suas características através de um hiperplano separador:



Espaço dos Insumos

Espaço das características

Voltar

SVM – Derivação no caso separável

No caso separável (HASTIE et al, 2009), temos que se há N pares

$$(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N) \text{ com } x_i \in R^N \text{ e } y_i \in \{-1, 1\}$$

Defina um hiperplano como: $\{x : f(x) = x'\beta + \beta_0 = 0\}$

Adicionalmente, defina M como a margem mínima desse hiperplano em relação ao ponto de treinamento. Então, o problema de encontrar o hiperplano separador é dado por:

$$max_{\beta, \beta_0, ||\beta||=1} M$$
$$\text{subject to } y_i(x_i'\beta + \beta_0) \geq M, i = 1, \ldots, N$$

O classificador é dado por $G(x) = sign[x'\beta + \beta_0]$

Voltar

Qual a necessidade do exercício de classificar imóveis?

- A necessidade de classificação dos imóveis advém da percepção que imóveis classificados atualmente como casas, galpões e barracões possam estar sendo valorados no mercado de terrenos e não no mercado próprio de sua tipologia.
- Assim, por exemplo, uma casa antiga num lugar valorizado da cidade pode ser negociada por um preço relativamente incondizente com suas características que usualmente são mais importantes: números de cômodos e acesso a amenidades, etc. Na verdade, a hipótese é de que se o preço seja incondizente com esse perfil dos imóveis, ela está sendo valorada por características usualmente mais importantes no mercado de lotes, tais como potencial construtivo, coeficiente de aproveitamento possível no local e possíveis empreendimentos valorizadores na vizinhança, tais como shoppings e vias importantes.
- Os imóveis a serem classificados são do tipo casa, barracões e galpões que podem estar sendo negociados na verdade como lotes. A ideia é que caso esses imóveis estejam associados a algum projeto, então é muito provável que eles estejam sendo negociados como lotes.

Qual a necessidade do exercício de classificar imóveis?

- Esse problema não existe com os imóveis já presentes na base de projetos, os quais a prefeitura já dispõe de informações sobre as mudanças estruturais e de tipologia. Contudo, é usual o descasamento dos imóveis presentes na base de dados do ITBI e aqueles que entram na base de dados do projeto.
- Na verdade, é provável que o imóvel seja negociado e precificado como lote, contudo a prefeitura só terá ciência dessa percepção do mercado quando o dono do imóvel a notificar oficialmente de mudanças na estrutura e tipologia do imóvel.
- Assim sendo, o modelo de classificação tem por objetivo se antever de certa forma a esse descasamento entre imóveis negociados no mercado imobiliário e conseguir estimar um preço mais condizente com a percepção que o mercado tem de um certo imóvel.
- O modelo de classificação assim receberia um imóvel do tipo casa, galpão e barracão e o classificaria com base no treinamento realizado em 2020, pelo CEDEPLAR, como sendo negociado no mercado de lotes ou em mercado próprio.
- Logo, o uso do modelo de classificação é um insumo para os modelos de preços.
- Analogamente ao modelo de preços, houve três modelos de classificação treinados: um para cada uma das seguintes tipologias de imóveis: casas, galpões e barracões.

Qual a necessidade do exercício de classificar imóveis?

Em termos práticos, o modelo se alimenta de um novo imóvel na base ITBI ou IPTU, com as mesmas variáveis as quais ele foi treinado e retorna um valor de 0 ou 1, com o 1 indicando que o imóvel foi negociado no mercado de lotes e 0, se foi negociado em mercado imobiliário próprio (casas sendo negociadas no mercado de casas e assim por diante).

O modelo pode ser atualizado a qualquer momento com novas entradas de novos imóveis no ITBI que podem ser conjugados ou não a algum projeto na Base de dados de Projetos da PBH.

Voltar

Divisão da Amostra

- Divisões usuais da amostra em treinamento e teste geralmente constem em 80/20 ou 75/25, isto é, 80% ou 75% da base para treinamento e 20% ou 25% da base para teste.
- Há *trade-offs* inerentes nessa divisão: quão maior o percentual usado na base de treinamento, maior a chance de o modelo falsamente acusar grande acerto *in-sample*, isto é, considerando a base total, treinamento e teste, contudo maior a chance de ele não generalizar bem *out-of-sample*, ou seja, para novos dados.
- Dois outros problemas relevantes numa tarefa de classificação dizem respeito ao *overfitting* e o *underfitting*.
- O *overfitting* ocorre quando se utiliza várias variáveis ou preditores na tarefa de classificação de tal forma que o modelo acusa um ajuste fino e preciso dos dados da amostra de teste, contudo ele falha para novos dados. Fraseando de outra forma, o modelo é um bom preditor *in-sample*, mas um péssimo preditor *out-of-sample*.
- O *underfitting* ocorre quando o modelo utilizado não aproveita a informação presente na base de treinamento com eficiência, sendo um modelo não interessante tanto *in-sample* quanto *out-of-sample*. Isso tende a ocorrer quando o número de vaariáveis é baixo.
- Esses dois problemas ocorrem em direções contrárias, ou seja, poucas variáveis é problemático assim como um número excessivo delas. Logo, um equilíbrio é necessário.
- Uma forma de contornar desse problema consiste numa validação cruzada, não realizada aqui. Mas em geral, invés de realizar o treinamento em apenas um subamostra da base original, realiza-se o mesmo exercício de classificação em várias subamostras aleatórias da base original e depois utiliza-se o melhor resultado médio delas. Isso tende a corrigir a chance de ocorrência dos dois problemas exibidos anteriormente.

Voltar

Implementação em outras linguagens

- O algoritmo do Gradient Boosting se encontra implementado em várias linguagens, tais como o R, Python, Matlab, Julia, dentre outras. A seguir, apresento o link da documentação nessas linguagens listadas:
- Python
- R
- Matlab
- Julia

Voltar

Detalhamento das variáveis: efeito esperado

Espera-se que as variáveis escolhidas tenham o seguinte impacto em termos de ajudar na classificação de um imóvel como sendo negociado no mercado de lotes ou em mercado imobiliário próprio:

1. Valor declarado do imóvel: em caso de imóvel localizado em área valorizada da cidade, isso deveria ser um bom preditor para que imóveis mais caros estão sendo negociados no mercado de lote. A interação dessa variável com outras características dos imóveis pode ajudar a decidir mais precisamente em qual mercado um imóvel está sendo negociado.
2. Ano de transação: imóveis negociados mais recentemente podem ter projetos associados de forma mais rápida.
3. Padrão de acabamento (sob a forma de cinco dummies): imóveis com alto padrão de acabamento correm menos risco de estarem sendo negociados no mercado de lotes.
4. Área construída adquirida: espera-se que a interação dessa variável com a área do terreno adquirida seja um bom preditor em termos da perda de construção que o proprietário tende a incorrer com a mudança da estrutura e tipologia do imóvel.
5. Área do terreno adquirida: vide item anterior.
6. Idade do imóvel: Imóveis mais velhos provavelmente estão sendo negociados no mercado de lotes e não no mercado da sua própria tipologia.
7. IQVU: esse índice pode ajudar na classificação de imóveis como sendo negociados no mercado de lotes, em especial no que concerne a imóveis em áreas com alta qualidade de vida, como medida pelo índice.

Detalhamento das variáveis: efeito esperado

Espera-se que as variáveis escolhidas tenham o seguinte impacto em termos de ajudar na classificação de um imóvel como sendo negociado no mercado de lotes ou em mercado imobiliário próprio:

1. Distância do imóvel ao hospital mais próximo: mensura acesso a amenidades. Essa variável interagida com outras variáveis podem contribuir para a melhor classificação do imóvel.
2. Distância do imóvel a praças regionais: vide item acima.
3. Média dos imóveis vizinhos que são apartamentos (wAP): essa informação revela duas informações; (i) indiretamente o potencial construtivo na vizinhança de um imóvel; (ii) pressão imobiliária no seu entorno. Logo, valores muito altos dela podem revelar maior probabilidade de o imóvel estar sendo negociado no mercado de lotes.
4. Média da idade dos imóveis vizinhos (widade): variável que mensura a pressão imobiliária sobre uma outra ótica, qual seja: empreendimentos imobiliários no entorno do imóvel. Quanto menor seu valor, mais aquecido o mercado no entorno do imóvel.
5. Interação entre wAP e widade: essa interação contribui por adicionar uma medição extra da pressão imobiliária no entorno do imóvel.
6. Dummy_app: dummy se o imóvel está ou não em área de proteção Ambiental: diz respeito ao acesso a amenidades no entorno do imóvel. Essa variável interagida com outras variáveis podem contribuir para a melhor classificação do imóvel.

Voltar