SMNE UFPR
Simpósio de Métodos Numéricos em Engenharia
25 a 27 de outubro, 2017

Aprendizagem de Máquina Aplicada a Investimentos Imobiliários

Natalha Cristina da Cruz Machado Benatti
Departamento de Matemática
Universidade Federal do Paraná
Curitiba, Brasil
natalha.benatti@ufpr.br

***Resumo*—Este trabalho apresenta estratégias utilizadas para classificar um imóvel em subvalorizado ou supervalorizado para a devida recomendação deste para compra como forma de investimento. Foi realizado um levantamento de dados de 150 apartamentos na região do bairro Água Verde, em Curitiba, PR, contando com 10 variáveis independentes e uma variável dependente. Foram utilizados os modelos de Regressão Logística e de Máquinas de Vetores Suporte com margens flexíveis abordada com e sem o uso de funções *kernel*. Os resultados foram comparados dentro desta problemática e então determinado o melhor classificador, resultando em um modelo de fácil aplicação para o utilizador.**

***Palavras-chave*—Avaliação de Imóveis, Aprendizagem de Máquina, Máquinas de Vetores Suporte, Regressão Logística.**

I. Introdução

Um imóvel é um patrimônio físico, um bem tangível já que pode ser tocado e visto, mas que não pode ser tomado de assalto como a maioria dos outros bens físicos existentes. O dinheiro investido em um imóvel não se perde com facilidade, por isso, dentre todos os tipos de investimento o imobiliário é um dos que possuem maior estabilidade. As quedas de preços também ocorrem no mercado imobiliário, mas não de forma tão brusca como em outras opções de investimento.

Imóveis podem ser transformados, se uma área que antes era residencial se tornar uma área comercial, o imóvel poderá ser modificado, tornando-se mais valioso, por exemplo. Mesmo com essa aparente facilidade relacionada aos investimentos em imóveis, é importante conhecer o máximo possível sobre o mercado. Quanto mais conhecimento o investidor obter sobre o ramo, maior será o potencial de seu ganho.

Há diversas maneiras de avaliar a rentabilidade futura de um investimento imobiliário, como o valor de seu aluguel ou o potencial de valorização do imóvel. Segundo [2] a possibilidade de obter bons lucros aumenta quando o investidor já encontra um imóvel que por alguma razão esteja subvalorizado para venda.

Muitos imóveis ficam subvalorizados quando há uma grande oferta de imóveis novos em uma cidade ou em uma certa região. Outros se desvalorizam por conta da proximidade de obras que, em algum momento, irão acabar. Outros, por ondas de furtos na região. Esta subvalorização também pode estar ligada à partilha dos bens de uma herança, a divisão dos bens relacionados a uma separação matrimonial, a deterioração do imóvel, dentre outros fatores.

A compra de um imóvel que está subvalorizado é no geral um bom investimento, exceto quando esta subvalorização está ligada à fatores permanentes. O lucro que provém da compra de imóveis subvalorizados exige muitas vezes paciência, de forma que o fator de desvalorização já não exista, ou pode depender diretamente do investidor, com pequenas reformas para revitalização do imóvel, quando esta desvalorização está ligada a deterioração do mesmo.

Todos precisam e sempre irão precisar de um imóvel para morar, por isso sempre existirá demanda por imóveis para comprar ou alugar.

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES.

Em alguns momentos esta demanda poderá ser menor ou maior, mas o mercado imobiliário oferece oportunidades em todo tempo, em determinados momentos oferece oportunidades de compra, em outros, de venda. Podemos ganhar dinheiro na compra e na venda e por isto, para o investidor, não importa se o mercado passa por uma crise, pois é quando se compra imóvel barato, ou se passa por um momento de euforia, pois é o momento de vender o imóvel mais caro.

Uma questão natural que se segue é, dado um imóvel e suas características, como avaliar se o imóvel a venda está ou não subvalorizado. Esta análise pode ser feita através da pesquisa de mercado, comparando o valor de imóveis com características semelhantes. Nosso trabalho visa a classificação de um imóvel como subvalorizado ou supervalorizado, estimulando a tomada de decisão de um investidor quanto à compra de dado imóvel como forma de investimento ou não. Esta classificação será feita utilizando métodos de aprendizagem de máquina, como a Regressão Logística e Máquinas de Vetores Suporte, que serão apresentadas a seguir. Vale salientar que para classificação de determinado imóvel, consideramos algumas características que influenciam diretamente em seu valor, além de seu valor de venda atual.

II. Levantamento de Dados

A obtenção de amostras válidas e confiáveis de dados imobiliários constitui uma das etapas mais importantes, e a que apresenta as maiores dificuldades, no processo avaliatório. Diversos fatores contribuem para isto, como a peculiaridade do mercado, as características específicas do imóvel avaliado, como sua dimensão, padrão, localização, entre outros.

Nosso banco de dados contém 150 amostras de imóveis, todas as amostras são representadas por apartamentos em condomínio fechado localizadas na região do bairro Água Verde, em Curitiba, Paraná.

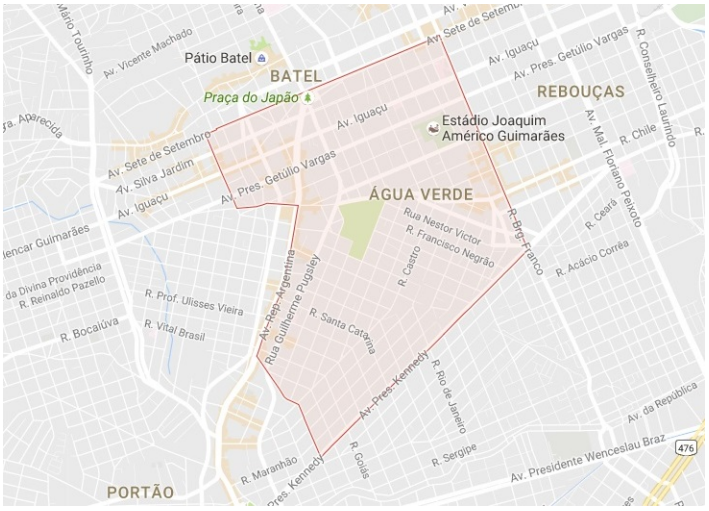

Figura 1: Delimitações do bairro Água Verde, Curitiba, Paraná.

Destas 150 amostras, 120 constituem o conjunto de treinamento do modelo e as 30 restantes o conjunto de teste. O conjunto de treinamento é utilizado para obtenção do classificador, enquanto que o conjunto de teste é utilizado para conhecer sua taxa de acerto.

Obtivemos tais dados a partir do site *Zap Imóveis* [8], sendo um dos maiores portais de classificados de imóveis no Brasil, acedido através de

www.zapimoveis.com.br.

Para avaliação de cada imóvel, consideramos as características:

- Número de Quartos - Variável contempla o número de quartos do imóvel, exceto suítes;
- Número de Suítes - Variável que contempla o número de suítes do imóvel;
- Número de Banheiros - Variável que considera o número de banheiros do imóvel, estando inclusos os banheiros sociais, os lavabos e os banheiros anexos às suítes;
- Área Útil da Unidade - É a área efetivamente utilizada do apartamento, incluindo paredes internas, sacadas e floreiras. É delimitada pelas paredes externas da fachada, pelas paredes da área de uso comum e pelas paredes de unidades autônomas. Tal variável é expressa em metros quadrados;
- Área Total da Unidade: É o somatório da área útil, área das vagas de garagem e da área comum do imóvel, não contemplando a área comum do empreendimento, como playground, salão de festas, quadras esportivas, piscinas, etc. Tal variável é expressa em metros quadrados;
- Vagas de Garagem: Variável discreta que contempla o número de vagas de garagem do imóvel;
- Número de Andares do Empreendimento: Variável discreta que considera o número de andares do empreendimento ao qual o imóvel se situa;
- Valor Médio do Condomínio: Contempla o valor médio da taxa de condomínio do imóvel. Tal valor é estimado para empreendimentos que ainda estão em processo de construção;
- Padrão do Empreendimento: Esta variável foi coletada através da observação da fachada do empreendimento, bem como o acabamento no interior do imóvel, utilizando as imagens fornecidas pelo site. Assume valores 1 - Alto padrão, 2 - Médio padrão e 3 - Baixo padrão;
- Valor de Venda: Variável que descreve o valor estipulado pelo proprietário para venda do imóvel. Vale salientar que tal variável foi coletada no dia 14/10/16 para todo o banco de dados.

Para determinar a subvalorização ou supervalorização do imóvel, fez-se necessário a utilização de um site de avaliação de imóveis, nomeadamente *AImóveis* [1], acedido através de

www.aimoveis.net

onde dadas as variáveis acima descritas (exceto o valor do imóvel), é calculada uma estimativa do valor de venda para o imóvel. Para

imóveis que estão com valores de venda menores que o valor estimado pelo site, o consideramos subvalorizado $(+1)$, onde a compra como forma de investimento será recomendada, caso contrário, se a valor de venda estiver maior ou igual o estimado o consideramos neutro ou supervalorizado (-1), não justificando o investimento.

- Avaliação do Investimento: Esta é a variável dependente do nosso modelo, considerando a subvalorização ou não do imóvel, tomando valores $+1$ ou -1 para uso de Máquinas de Vetores Suporte e $+1$ ou 0 para Regressão Logística.

III. Modelos Utilizados

Para criação do melhor modelo de classificação para o problema exposto utilizamos Regressão Logística e Máquinas de Vetores Suporte (SVM, do inglês *Support Vector Machines*). Apresentemos a seguir detalhes acerca dos métodos utilizados.

A. Regressão Logística

Regressão Logística é um modelo de classificação binária ou multinomial. Mais detalhes podem ser encontrados no livro de Scott Menard [4]. Nossa problemática se enquadra em classificação binária, onde naturalmente se considera uma variável dependente $y_i \in \{0,1\}$ indicando a classe a que o ponto $x^{(i)}$ pertence. Nesse contexto, X^- denota a classe negativa, isto é, os elementos x_i associados a $y_i = 0$, e X^+ denota a classe positiva, isto é, os elementos x_i associados a $y_i = 1$.

Dado o conjunto de dados $\{(x^{(1)}, y_1), (x^{(2)}, y_2), \ldots, (x^{(m)}, y_m)\} \subseteq \Omega \times \{0,1\}$, onde $\Omega = \{x^{(i)} = (1, x_1^{(i)}, \ldots, x_n^{(i)})\} \subseteq \mathbb{R}^{n+1}$, o modelo de Regressão Logística é dado por $m_\theta(x) = g(\theta^T x)$ com $g : \mathbb{R} \to [0,1]$ definida como $g(z) = \dfrac{1}{1+e^{-z}}$, onde $\theta = (\theta_0, \theta_1, \ldots, \theta_n)$ é minimizador da seguinte função custo

$$h(\theta) = -\sum_{i=1}^{m} \left[y_i \log\left(m_\theta(x^{(i)})\right) + (1-y_i)\log\left(1 - m_\theta(x^{(i)})\right)\right],$$

com $\log(t)$ denotando o logaritmo na base Neperiana. Desta forma, introduzido um novo dado x, tal é classificado como pertencente a classe positiva X^+ caso $m_\theta(x) \geq 0.5$ e pertencente a classe negativa X^- caso $m_\theta(x) < 0.5$. Para resolvermos o problema de Otimização

$$\min_{\theta \in \mathbb{R}^{n+1}} \quad h(\theta) \tag{1}$$

utilizamos a função *fminunc* do *software* MATLAB e o Método do Gradiente Acelerado de Nesterov [5], ambos tomando a origem como ponto inicial.

B. Máquinas de Vetores Suporte

Fundamentado na Teoria da Aprendizagem Estatística, o método de classificação de dados usando máquinas de vetores suporte [7] é objeto de pesquisa desde a década de 60. As máquinas de vetores suporte para classificação têm o intuito de classificar objetos n-dimensionais $x = (x_1, x_2, \ldots, x_n)^T$ utilizando, para isso, um hiperplano. No caso mais simples, consideramos duas classes, às quais chamamos classe positiva X^+ e negativa X^-. Para encontrar o padrão dos elementos analisados são utilizadas amostras onde já se conhece a classe a que cada elemento do conjunto de dados pertence, criando um modelo para futuras avaliações.

Consideremos o conjunto de dados $\{(x^{(1)}, y_1), (x^{(2)}, y_2), \ldots, (x^{(m)}, y_m)\} \subseteq \Omega \times \{+1, -1\}$, com $\Omega = \{x^{(i)} = (x_1^{(i)}, \ldots, x_n^{(i)})\} \subseteq \mathbb{R}^n$ a união disjunta dos conjuntos X^+ e X^-, onde $y_i = 1$ quando $x^{(i)} \in X^+$ e $y_i = -1$ quando $x^{(i)} \in X^-$. O objetivo é encontrar um hiperplano ${w^*}^T x + b^* = 0$ que separe os pontos de classes distintas, definindo o modelo dado por $m(x) = \text{sinal}(\theta^T x + b)$ onde (θ, b) é a solução do problema de Otimização expresso a seguir.

$$\begin{array}{lll} \min\limits_{\theta \in \mathbb{R}^n, b \in \mathbb{R}, \xi \in \mathbb{R}^m} & \frac{1}{2}\|\theta\|_2^2 + C\sum\limits_{i=1}^{m} \xi_i & \\ \text{s.a} & y_i(\theta^T x^{(i)} + b) \geq 1 - \xi_i & i = 1, \ldots, m \\ & \xi_i \geq 0 & i = 1, \ldots, m. \end{array} \tag{2}$$

Onde $C > 0$ é o parâmetro de regularização, e ξ_i são as variáveis de folga. O parâmetro C pondera a maximização da margem e o número de erros permitidos no conjunto de treinamento.

Geralmente, o problema (2) é resolvido em sua formulação dual, pois este nos dá como informação a taxa de vetores suporte do classificador, sendo tais os dados $x^{(i)}$ correspondentes aos α_i não nulos. O conhecimento desta taxa de vetores suporte é de grande importância na teoria de Máquinas de Vetores Suporte, e mais detalhes podem ser encontrados em [3]. A formulação dual do problema (2) é dada por

$$\begin{array}{lll} \max\limits_{\alpha \in \mathbb{R}^m} & -\frac{1}{2}\sum\limits_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j {x^{(i)}}^T x^{(j)} + \sum\limits_{i=1}^{m} \alpha_i & \\ \text{s. a} & \sum\limits_{i=1}^{m} \alpha_i y_i = 0 & i = 1, \ldots, m \\ & 0 \leq \alpha \leq C. & \end{array} \tag{3}$$

Mediante ao avanço da teoria de Máquinas de Vetores Suporte, notou-se que há casos onde o classificador apresentado a priori não traz resultados satisfatórios. Isto motivou a utilização de funções *kernel*, onde consideramos $\varphi : \mathbb{R}^n \to \mathbb{R}^S$ uma função de mapeamento, que parte do espaço de origem e vai para o espaço de características que se dá em uma dimensão maior. Assim o problema de minimização é dado por:

$$\begin{array}{lll} \min\limits_{\theta \in \mathbb{R}^S, b \in \mathbb{R}, \xi \in \mathbb{R}^m} & \frac{1}{2}\|\theta\|_2^2 + C\sum\limits_{i=1}^{m} \xi_i & \\ \text{sujeito a} & y_i(\theta^T \varphi(x^{(i)}) + b) \geq 1 - \xi_i & i = 1, \ldots, m \\ & \xi_i \geq 0 & i = 1, \ldots, m. \end{array} \tag{4}$$

A formulação do problema dual é dado por

$$\begin{aligned} \max_{\alpha\in\mathbb{R}^m} \quad & \sum_{i=1}^{m}\alpha_i - \tfrac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j \varphi(x^{(i)})^T\varphi(x^{(j)}) \\ \text{s. a} \quad & \sum_{i=1}^{m}\alpha_i y_i = 0 \qquad\qquad i=1,\dots,m \\ & 0\le\alpha\le C, \end{aligned} \tag{5}$$

e mais detalhes podem ser vistos em [6]. Nesse caso, não é necessário o conhecimento explicito da função de mapeamento ϕ, mas apenas do produto interno $\phi(x^{(i)})^T\phi(x^{(j)})$. Assim, o que se faz na prática é tomar uma função $K:\mathbb{R}^n\times\mathbb{R}^n\to\mathbb{R}$ de maneira que exista um mapeamento ϕ com $K(x^{(i)},x^{(j)}) = \varphi(x^{(i)})^T\varphi(x^{(j)})$. As funções K que cumprem esta condição são denominadas funções *kernel*. Desta forma, o problema (5) pode ser reescrito como

$$\begin{aligned} \max_{\alpha\in\mathbb{R}^m} \quad & \sum_{i=1}^{m}\alpha_i - \tfrac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j K(x^{(i)},x^{(j)}) \\ \text{s. a} \quad & \sum_{i=1}^{m}\alpha_i y_i = 0 \qquad\qquad i=1,\dots,m \\ & 0\le\alpha\le C. \end{aligned} \tag{6}$$

Tal problema é dado na seguinte forma matricial

$$\begin{aligned} \max_{\alpha\in\mathbb{R}^m} \quad & -\tfrac{1}{2}\alpha^T H\alpha + e^T\alpha \\ \text{s.a} \quad & y^T\alpha = 0 \\ & 0\le\alpha\le C \end{aligned} \tag{7}$$

com $e_j = 1$, $i,j = 1,\dots,m$, e $H_{ij} = y_i y_j K(x^{(i)},x^{(j)})$.

Em nossa implementação foram utilizadas as seguintes funções *kernel*:

- Kernel Gaussiano: $K(x^{(i)},x^{(j)}) = \exp\left(-\gamma\|x^{(i)}-x^{(j)}\|^2\right)$, com $\gamma>0$;
- Kernel Sigmoidal: $K(x^{(i)},x^{(j)}) = \tanh(\gamma({x^{(i)}}^T x^{(j)}) + \eta)$, com $\gamma>0$;
- Kernel Polinomial de grau p: $K(x^{(i)},x^{(j)}) = \left({x^{(i)}}^T x^{(j)} + \eta\right)^p$.

Os problemas do âmbito de SVM foram resolvidos através da função *quadprog* do *software* MATLAB.

IV. Experimentos Numéricos

Os testes foram realizados em um computador portátil com processador AMD A8-5550M APU, velocidade do clock de 2.10 GHz, 4 GB de memória RAM e sistema operacional Windows 8.1 Pro com arquitetura 64 bits. Os algoritmos foram implementados no software Matlab versão 8.3 (R2014a). O objetivo aqui foi encontrar o melhor classificador para os dados deste problema específico.

A. Resultados Numéricos

Primeiramente, reordenamos de maneira aleatória nosso conjunto de dados através de um vetor gerado pelo comando *randperm* fixada a semente de aleatoriedade para todas as implementações. Utilizamos a metodologia de validação cruzada do tipo *K-fold* que consiste em dividir o conjunto de amostras em K subconjuntos de mesmo tamanho, treinar o algoritmo com $K-1$ subconjuntos e testá-lo com o remanescente. Repetindo este processo K vezes para garantir que todos os subconjuntos sejam avaliados, contabilizando-se a taxa de acerto em cada processamento. Em nossa implementação tomamos $K = 5$, onde cada subconjunto tem 30 amostras, obtendo para cada processamento o conjunto de teste localizado entre as linhas $1+30(t-1)$ e $30t$ da matriz de dados, para $t = 1,\dots,K$.

Em nosso primeiro teste, utilizamos o Modelo de Regressão Logística, resolvendo o problema (1) utilizando a função Fminunc, própria do Matlab, e o Gradiente Acelerado de Nesterov. Relativamente ao Modelo C-SVM, para realizar a escolha de C utilizamos a clássica busca por *Grid*, tomando assim $C = 2^j$, com $j\in\{-8,-7,\dots,7,8\}$. Vale ressaltar que há outras formas de obtenção dos parâmetros associados às Máquinas de Vetores Suporte além da busca por *Grid*, como pode-se ver em [3]. A priori, resolvemos o problema dual (3), isto é, sem o uso de funções *kernel*. Posteriormente, resolvemos o problema (6), que equivale a (7), utilizando os *kernels* Gaussiano, Sigmoidal e Polinomial. Os resultados dos testes podem são apresentados na tabela a seguir:

Método Utilizado	**Taxa de Acerto**
Regressão Logística (fminunc)	50%
Regressão Logística (Gradiente Acelerado de Nesterov)	50%
C-SVM (sem utilização de Kernel)	86.67%
C-SVM (com Kernel Gaussiano)	83.33%
C-SVM (com Kernel Sigmoidal)	83.33%
C-SVM (com Kernel Polinomial)	85.56%

Tabela I: Resultados numéricos, expressando o método utilizado e a taxa de acerto.

Portanto, o modelo que obteve melhores resultados quando testado no conjunto de teste foi C-SVM sem uso de funções *kernel*. Veremos a seguir um caso particular de uma avaliação realizada com este classificador.

B. Aplicação do Modelo em um Caso Particular

No Edifício Prime Class, em Outubro de 2016, havia um apartamento a venda com as seguintes características.

- 2 quartos;

- 1 suíte;
- 2 banheiros;
- 120 m^2 de área útil;
- 176 m^2 de área total;
- 1 vaga de garagem;
- Edifício de 8 andares;
- Média de R$600 de taxa de condomínio;
- Padrão médio;
- Valor de venda: R$ 760.000.

Tomando então

$$x = [2,\ 1,\ 2,\ 120,\ 176,\ 1,\ 8,\ 600,\ 2,\ 760000]^T,$$

e o modelo obtido para avaliação, temos

$$\begin{aligned} \text{sinal}({\theta^*}^T x + b^*) &= \text{sinal}(0.4499) \\ &= 1 \end{aligned}$$

Isto é, o imóvel foi avaliado como subvalorizado, sendo assim recomendado como uma opção para investimento. Note que tal classificação foi condizente com a feita pelo site *AImóveis*, que avaliou o imóvel segundo suas caraterísticas por R$ 810.000, caracterizando-o como subvalorizado.

V. Conclusão e Trabalhos Futuros

Como vimos nos experimentos numéricos, o modelo que mostrou melhor eficiência para o nosso problema foi C-SVM sem uso da função *kernel*, obtendo a melhor taxa de acerto no conjunto de testes, $86,67\%$. Tal classificador é definido por $m(x) = \text{sinal}({\theta^*}^T x + b^*)$, com conjuntos de treinamento e teste definidos por $t = 5$, e pelo uso da constante de regularização $C = 8$. Assim, é possível classificar um novo imóvel como subvalorizado ou supervalorizado, consoante às suas características, para a devida recomendação de compra como forma de investimento para o usuário. De certa forma, este classificador ausenta a necessidade de contratação um corretor de imóveis para avaliação de tais.

Ideias para trabalhos futuros contemplam a expansão do banco de dados para amostras de imóveis em outros bairros de Curitiba ou ainda outras cidades. Além disso, a seleção de variáveis com maior impacto no modelo, isto é, a redução do número de características, considerando somente as que tem mais importância na construção da máquina.

Agradecimentos

Gostaria de agradecer primeiramente à CAPES, por ter me financiado durante todo o percorrer do Mestrado em Matemática na UFPR. E quero agradecer também ao Dr. Geovani Grapiglia, professor do Departamento de Matemática da UFPR, por ter incentivado e me auxiliado a dar vida a este trabalho.

Referências

[1] Aimoveis. Disponível em: <www.aimoveis.net/public/appraisals> acedido em 16 de Setembro de 2016.
[2] D. M. S. Andrade. *O Melhor Investimento para Você: Princípios de Educação Financeira*. 1 ed. Rio de Janeiro. AR Editora, 2015.
[3] N. C. C. M. Benatti. *Métodos de Busca Direta para Seleção de Parâmetros em Máquinas de Vetores Suporte.* Dissertação de Mestrado, Universidade Federal do Paraná, Curitiba, 2017.
[4] S. Menard. *Applied Logistic Regression Analysis.* Vol. 106. Sage, 2002.
[5] Y. Nesterov. *Gradient Methods for Minimizing Composite Objective Function.* CORE: Discussion. Papers 76, 2007.
[6] A. Shmilovici. Support Vector Machines. In: Maimon, O.; Rokach, L. *Data Mining and Knowledge Discovery Handbook*. Springer, London, 2010. P.231-235.
[7] V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer Verlag, New York, 1995.
[8] Zap Imóveis. Disponível em: <www.zapimoveis.com.br> Acesso em 10 de Setembro de 2016.