

FUNDAÇÃO GETULIO VARGAS
ESCOLA DE MATEMÁTICA APLICADA – FGV/EMAp
CURSO DE GRADUAÇÃO EM MATEMÁTICA APLICADA

PREVISÃO DE RESULTADOS DE TÊNIS.

BERNARDO RUSSO GUEDES FERREIRA

Rio de Janeiro
2018

FUNDAÇÃO GETULIO VARGAS
ESCOLA DE MATEMÁTICA APLICADA – FGV/EMAp
CURSO DE GRADUAÇÃO EM MATEMÁTICA APLICADA

PREVISÃO DE RESULTADOS DE TÊNIS.

“Declaro ser o único autor do presente projeto de monografia que se refere ao plano de trabalho a ser executado para continuidade da monografia e ressalto que não recorri a qualquer forma de colaboração ou auxílio de terceiros para realizá-lo a não ser nos casos e para os fins autorizados pelo professor orientador.”

Bernardo Russo Guedes Ferreira

Orientador: Moacyr Alvim Horta B. da Silva

Rio de Janeiro
2018

BERNARDO RUSSO GUEDES FERREIRA

PREVISÃO DE RESULTADOS DE TÊNIS.

“Trabalho de Conclusão apresentado à Escola de Matemática Aplicada como requisito para a obtenção parcial do grau de bacharel em Matemática Aplicada.”

Aprovado em ____ de ____ de ____.

Grau atribuído ao Trabalho de Conclusão: ____.

Professor Orientador: Moacyr Alvim Horta B. da Silva
Escola de Matemática Aplicada - FGV/EMAp
Fundação Getúlio Vargas

Sumário

1	Introdução	4
2	Pesquisa exploratória	5
3	Os Modelos	6
4	Os Dados	10
5	Resultados do Modelos	12
6	Estratégias de Apostas	14
7	Trabalhos Futuros	16

1 Introdução

Motivação

Esportes são tema de grande interesse nas horas de lazer e também são responsáveis por grandes movimentações financeiras, de modo que ser capaz de modelar e fazer palpites educados sobre futuros resultados em algum esporte é uma habilidade atraente.

Escolhemos o tênis como objeto do nosso estudo porque acreditamos que a sua natureza de enfrentamentos individuais poderia ser um fator de influência positiva na nossa capacidade de prever o vencedor. Além disso, temos disponíveis dados de todas as partidas da ATP desde 1968 no repositório do GitHub de Jeff Sackmann.

Resumo

Começamos este trabalho com uma pesquisa exploratória dos modelos atualmente usados na previsão de resultados no esporte, levando a uma comparação de suas acurácias na base de dados histórica de partidas. Determinado o modelo de melhor performance, atacamos a possibilidade de ganhos em apostas para o US Open 2018.

2 Pesquisa exploratória

Para a elaboração do trabalho, buscamos encontrar os artigos mais recentes que trouxessem modelos preditivos para esportes (em particular o tênis) juntamente com suas acurácias e, quando possível, performance contra casas de aposta.

Dentre os artigos lidos é válido citar alguns dos que tiveram maior influência sobre trabalho final.

[1] No artigo de Stanford, os autores descrevem alguns modelos de Machine Learning e aplicam aos mesmo dados a que temos acesso. Entre SVM, Redes Neurais, Random Forest e Regressão Logística, o artigo explica como foram construídas as features e apresenta os resultados dos algoritmos e retornos em apostas.

[2] Os autores deste artigo descrevem o modelo de Bradley-Terry, que é da seguinte forma: dados dois jogadores A e B , e suas proficiências a e b (cuja maneira de estimar varia entre autores), a probabilidade de vitória de A é dada por $\frac{a^e}{a^e + b^e}$ para um e a ser estimado no fit do modelo. O artigo explica uma maneira de se calcular as proficiências dos jogadores dados seus resultados anteriores e indica como atualizá-las com novas entradas.

[3] Este segundo é um post no blog de Jeff Sackman, Tennis Abstract, de autoria de Peter Wetz, onde ele compara alguns diferentes modelos estatísticos. Vale citá-lo porque o autor apresenta o score de Brier e confiabilidade dos preditores. Um dos modelos, o de Bradley-Terry, descrito no artigo anterior.

[4] Neste artigo, os autores analisam diversos casos do uso de redes neurais na previsão de resultados em esportes. Também é apresentado o sport result prediction CRISP-DM (SRP-CRISP-DM), que é uma adaptação da pipeline usual em projetos de data-mining, para o caso da previsão de resultados esportivos.

3 Os Modelos

Nesta seção descrevo brevemente os modelos que, na literatura, obtiveram os melhor resultados na previsão de partidas.

Regressão Logística

O modelo logístico trabalha com dados binários, nos quais o evento tem duas possibilidades: acontece (1) ou não acontece (0). Então, dado um conjunto de informações x , tentamos determinar se o y correspondente é 1 ou 0. A natureza binária do modelo é interessante para descrever, numa partida de tênis entre dois jogadores A e B , a variável 'vitória do jogador A ', por exemplo.

Na Regressão Logística temos uma hipótese da forma:

$$h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$$

Assumimos que as variáveis y de classificação binárias são extraídas de uma distribuição tal que:

$$P(y = 1|x; \theta) = h_{\theta}(x)$$

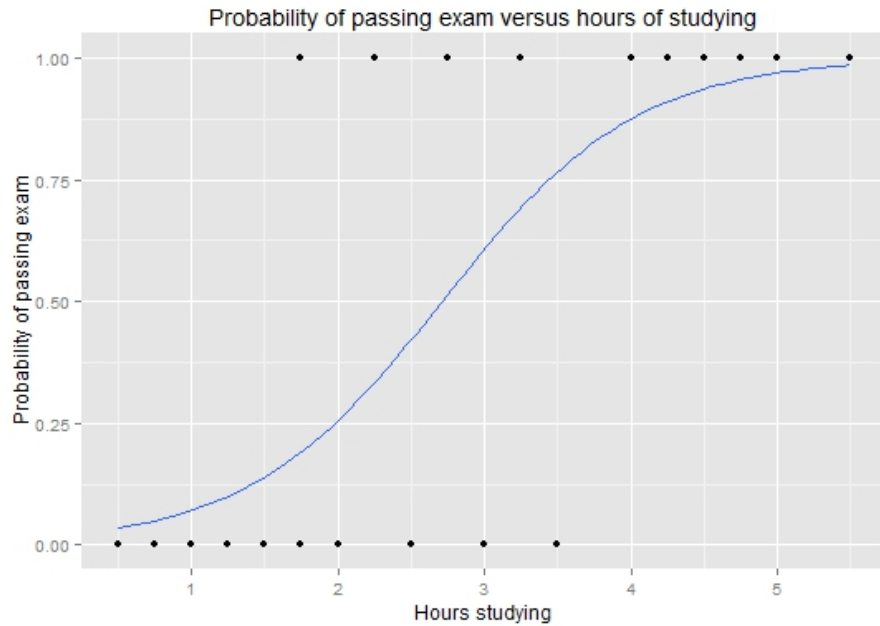
$$P(y = 0|x; \theta) = 1 - h_{\theta}(x)$$

Assim, dado um conjunto de partidas e seus resultados, nós escolhemos o θ que maximize a função de log-verossimilhança abaixo:

$$l(\theta) = \sum_i y_i \log(h_{\theta}(x_i)) + (1 - y_i) \log(1 - h_{\theta}(x_i))$$

Então, para uma nova entrada (partida) com parâmetros \hat{x} , dizemos que o \hat{y} correspondente é 1 se $h_{\theta}(\hat{x})$ e 0 caso contrário.

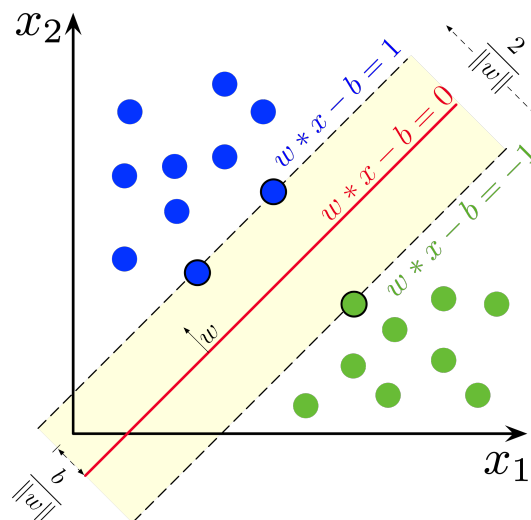
Abaixo, um exemplo do fit de uma curva logística em um conjunto de dados:



SVM

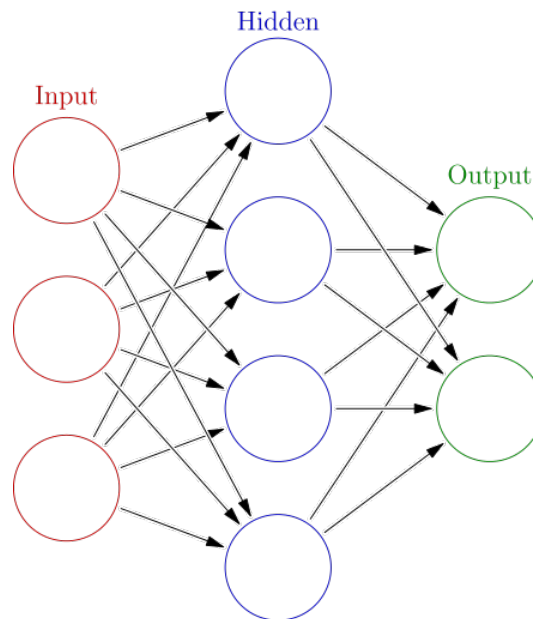
O algoritmo de SVM é um método de aprendizado supervisionado que separa os dados em categorias. Ele é treinado com dados pré categorizados e gera hiperplanos que separem esse dados nas suas respectivas categorias com a maior folga possível. Uma nova entrada x é então mapeada para esse espaço e, dependendo da posição da sua projeção em relação aos cortes, é imputada ao y correspondente uma categoria ou outra.

Os cortes podem ser tanto lineares quanto polinomiais. Abaixo, um exemplo de SVM linear para um problema com duas classes e entrada de dimensão 2.



Redes Neurais

Com sua arquitetura inspirada pelas redes de neurônios no cérebro, o modelo de Redes Neurais é implementado através de camadas de decisão, nas quais a informação é ponderada, transformada e repassada a outra camada. Esse processo é repetido até o momento que um neurônio de saída é ativado. Nesse momento é associada uma determinada categoria à entrada.

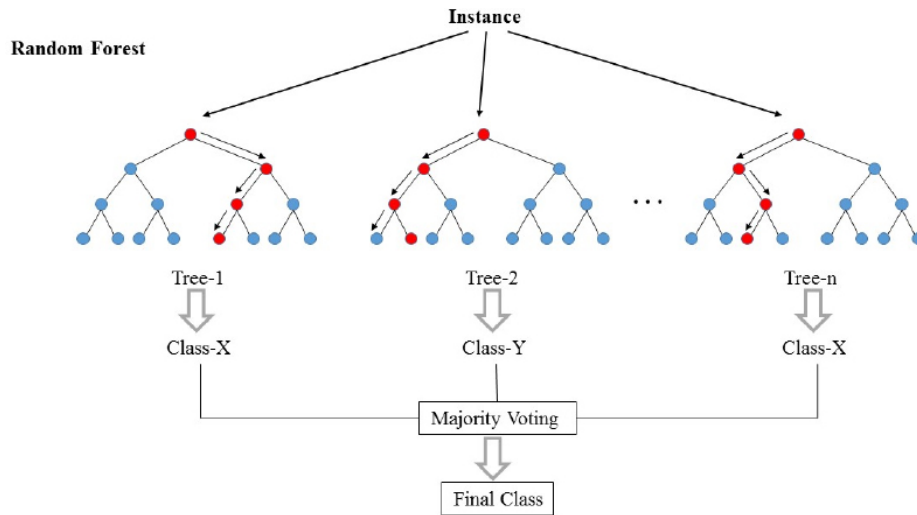


Random Forest

Uma árvore de decisão é um modelo preditivo no qual nós são usados para diferenciar, a partir do valor de uma determinada variável, se a entrada pertence a uma ou outra categoria.

O algoritmo Random Forest é um método de aprendizado que pode ser usado para classificações, regressões e outras tarefas e opera através da construção de diversas árvores de decisão. No caso de classificação, dados os outputs das diversas árvores, a classe mais votada é a classe final prevista pela floresta.

Abaixo, um exemplo de floresta com n árvores.



Bradley-Terry

O Bradley-Terry é um modelo de probabilidade usado para prever o resultado de uma comparação entre dois indivíduos. Dados dois jogadores A e B da população, com a e b valores reais positivos associados à habilidade dos indivíduos, o modelo estima a probabilidade

$$P(A > B) = \frac{a^e}{a^e + b^e}$$

A comparação $A > B$ pode ser lida, nesse contexto, como A ganha de B .

4 Os Dados

Os dados utilizados para treinar e testar os modelos são os encontrados no repositório do GitHub de Jeff Sackmann (https://github.com/JeffSackmann/tennis_atp), que abrange todas as partidas da ATP desde o seu início, em 1968. Cada entrada contém informações sobre ambos jogadores participantes (rank na ATP, mão boa, altura, idade...), a superfície da quadra (barro, grama ou cimento) e estatísticas da partida tais como número de aces, saques bloqueados, entre outros de cada jogador.

	tourney_id	tourney_name	surface	draw_size	tourney_level	tourney_date	match_num	winner_id	winner_seed	winner_entry	...	w_bpFaced	l_ace	l_df
0	2001-580	Australian Open	Hard	128.0	G	20010115.0	1.0	102856.0	1.0	NaN	...	17.0	8.0	6.0
1	2001-580	Australian Open	Hard	128.0	G	20010115.0	2.0	102257.0	NaN	NaN	...	10.0	1.0	6.0
2	2001-580	Australian Open	Hard	128.0	G	20010115.0	3.0	102905.0	NaN	NaN	...	10.0	10.0	3.0
3	2001-580	Australian Open	Hard	128.0	G	20010115.0	4.0	102694.0	NaN	Q	...	15.0	7.0	11.0
4	2001-580	Australian Open	Hard	128.0	G	20010115.0	5.0	102765.0	NaN	NaN	...	15.0	8.0	1.0

A base com os dados das casas de aposta está disponível gratuitamente no site <http://www.tennis-data.co.uk> e contém os odds realizados em diversas casas para partidas da ATP desde o ano de 2001.

	ATP	AvgL	AvgW	B&WL	B&WW	B365L	B365W	Best of	CBL	CBW	...	UBW	W1	W2	W3	W4	W5	WPts	WRank	Winner	Wsets
2111	53	18.45	1.01	NaN	NaN	23.0	1.01	5	NaN	NaN	...	NaN	6.0	3	NaN	NaN	NaN	10040.0	1.0	Nadal R.	1.0
2159	53	17.88	1.02	NaN	NaN	19.0	1.02	5	NaN	NaN	...	NaN	6.0	6	6	NaN	NaN	10040.0	1.0	Nadal R.	3.0
2178	53	8.56	1.08	NaN	NaN	9.0	1.08	5	NaN	NaN	...	NaN	5.0	7	7	7.0	NaN	10040.0	1.0	Nadal R.	3.0
2193	53	12.98	1.04	NaN	NaN	13.0	1.04	5	NaN	NaN	...	NaN	6.0	6	6	6.0	NaN	10040.0	1.0	Nadal R.	3.0
2201	53	4.04	1.26	NaN	NaN	4.0	1.25	5	NaN	NaN	...	NaN	0.0	6	7	6.0	7.0	10040.0	1.0	Nadal R.	3.0

Construção das Features

As features para os modelos preditivos foram construídas usando as variáveis que achamos que teriam influência sobre o resultado das partidas. O único preditor que teve um conjunto de features diferente dos demais foi o modelo de Bradley-Terry, no qual as únicas variáveis usadas na predição são as pontuações dos dois jogadores no ranking da ATP.

Para os demais modelos, as features foram selecionadas de maneira que não incluíssem informações sobre a partida a ser disputada ou quaisquer partidas futuras. Por motivos de simetria, todas as variáveis do jogo entre dois jogadores $p1$ e $p2$ são

da forma:

$$X_i = STAT_{i,p1} - STAT_{i,p2}$$

Algumas das estatísticas utilizadas são:

- Ranking na ATP;
- Pontuação no ranking;
- Média de estatísticas das últimas 5, 10 e 15 partidas (número de aces, pontos ganhos...)

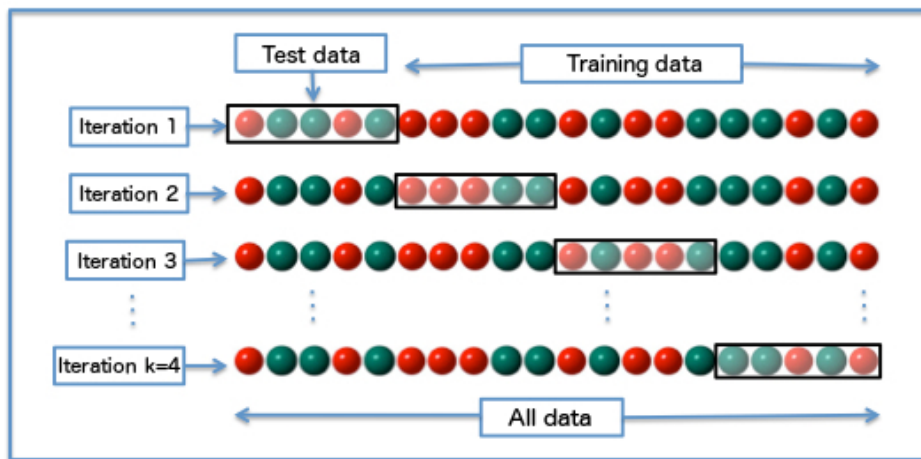
No fim do processo de construção, obtivemos um total de 45 variáveis, que foram depois normalizadas para o processo de treinamento e teste dos modelo.

5 Resultados do Modelos

Para comparação dos modelos, usamos suas versões do pacote Scikit-learn, implementadas em Python e testadas pelos diversos usuários do pacote. Comparamos os modelos usando um Cross-Validation, que é um dos métodos mais utilizados nesse tipo de situação, com $k = 5$ partições.

Cross-Validation

O Cross-Validation, ou validação cruzada, é feita particionando os dados em k conjuntos disjuntos e usando, em cada iteração $1 \leq i \leq k$, o i -ésimo conjunto como de teste e os demais para treino. Abaixo segue um esquema ilustrando uma aplicação para $k = 4$ partições.



A acurácia é então dada, para cada modelo, como a média das acurácias para cada uma das k partições. Assim, seguem abaixo os valores encontrados para a validação cruzada dos modelos.

Modelo		CV (k=5)
Regressão Logística	L2	66,3%
	L1	66,2%
Random Forest		65,3%
Redes Neurais		65,7%
SVM	Linear	66,1%
	Polynomial degree 3	65,8%
Bradley-Terry		62,7%

Não identificamos grande diferença de tempo para treino e teste entre os modelos de Regressão Logística, Random Forest e Redes Neurais. O de Bradley-Terry foi o mais rápido mas teve performance muito abaixo dos demais. Já o de SVM, mesmo não tendo uma alta acurácia, foi o que tomou mais tempo nas validações.

Comparando as métricas entre os modelos, decidimos usar, para estimação dos resultados, o de Regressão Logística com regularização L2.

6 Estratégias de Apostas

Nesta seção descrevemos a estratégia montada para definir se fazemos ou não uma aposta de acordo com as probabilidades dadas pelo modelo de predição e os odds da casa de apostas.

Os odds das casas de apostas são dados na forma decimal, ou seja, numa partida entre A e B , se a casa paga 2,75 e 1,40 para suas vitórias respectivamente, significa que as probabilidades implícitas de vitória dos dois são $P_A = \frac{1}{2,75} \approx 36,4\%$ e $P_B = \frac{1}{1,40} \approx 71,4\%$. Mas essas probabilidades, que são de um caso real, somam 107,8%, que é maior que os 100% que deveriam somar. Isso acontece porque a casa de aposta se utiliza do chamado 'overround', que é uma margem para que, na média, a casa tenha lucro.

Sabendo isso, só vale a pena apostar num determinado resultado se a probabilidade dada pelo nosso modelo para ocorrência daquele evento seja maior que a implícita paga pela casa.

No US Open de 2018, na partida entre Andreas Seppi e Sam Querrey, por exemplo, o nosso modelo apontava uma probabilidade de vitória de 43% de Seppi, bem maior que os 29% implícitos na casa Bet 365. Nesse caso, uma aposta no Seppi, apesar de não ser o jogador dado como favorito pelo modelo, geraria um retorno financeiro esperado positivo.

Assim, a estratégia ficou determinada da seguinte maneira: uma aposta é feita no jogador A sempre que a probabilidade de vitória dele seja maior que a implícita. Sejam P_A a probabilidade de vitória do jogador A em uma determinada partida e R_A o retorno dado pela casa no caso de vitória, o retorno sobre uma aposta de $R\$10$ no jogador A é $R\$(10R_A)$, caso ele vença e $-R\$10$ caso ele perca.

Na partida entre Seppi e Querrey, a nossa estratégia indicaria um potencial de ganho na aposta em Seppi e, de fato, teríamos um retorno positivo de 3,5 vezes o valor apostado.

O US Open 2018

Para prever os jogos do US Open 2018 decidimos usar o modelo de Regressão Logística com a regularização L2, por ser o que apresentou melhores resultados na

validação.

Para esse Grand Slam específico obtivemos um percentual de acerto de 76% das partidas. O resultado financeiro obtido foi um retorno médio de pouco mais de 4%.

A princípio este retorno parece promissor, mas devemos atentar ao fato que a acurácia de 76% nesse conjunto de dados foge bastante dos pouco mais de 66% no Cross-Validation. Podemos esperar que, expandindo o número apostas, o acerto convirja para os 66% e os retornos caiam.

7 Trabalhos Futuros

Em trabalhos futuros pretendo desenvolver tanto ideias que foram contempladas no projeto inicial mas não foram abordadas neste trabalho, quanto outras que surgiram durante o processo de produção deste. Abaixo algumas ideias que podem vir a ser desenvolvidas:

- Melhoria nos modelos:

Um ponto a ser atacado é aprofundar a nossa busca por modelos que possam prever bem os resultados das partidas. Adicionar também novas estatísticas sobre as partidas e refazer a construção das features são medidas que podem vir a trazer melhoras na habilidade de predição. A performance dos jogadores varia entre os tipos de quadra, integrar uma variável que explique isso no futuro pode trazer ganhos.

- Modelo de Oponentes comuns:

Uma das ideias iniciais era desenvolver um modelo de oponentes comuns que, além de auxiliar a prever resultados, de alguma forma nos desse um insight em como comparar a proficiência de jogadores e determinar a dimensão dessa. Podemos retomar isso em um próximo projeto.

- Ranking histórico dos jogadores:

Presente no projeto, gostaríamos de estudar, usando um modelo de oponentes comuns, como jogadores de épocas distintas podem ser comparados usando as proficiências vindas do modelo descrito acima.

- Consistência dos modelos:

Algo que não foi feito neste projeto e cuja realização seria de grande ganho para entender melhor os modelos é uma análise da consistência das previsões. Existem diversos tipos de maneiras de avaliar essa métrica em modelos e eu diria que este é o próximo passo mais natural.

- Modelo ELO:

Entre os modelos apresentados no artigo [3], está o de Bradley-Terry com as proficiências dos jogadores dadas em função do ranking ELO. Já difundido

em outros esportes tais como xadrez, tênis de mesa e baseball, o ELO é um método de pontuação que é atualizado proporcionalmente à probabilidade de o jogador vencer a partida.

- Revisão da estratégia de apostas:

Por não ser o principal foco do trabalho, optamos por uma estratégia de apostas simples, que explora as diferenças de julgamento do modelo e da casa de apostas. Um trabalho futuro pode vir a estudar como obter retornos melhores.

- Expansão para outros esportes:

Temos a ambição de, no horizonte, sermos capazes de prever resultados nos mais diversos esportes, por isso este item entra nessa lista. Fecho a lista com este item porque acho que ele vem, em prioridade, de fato depois dos anteriores.

References

- [1] G. S. Andre Cornman and D. Wright, “Machine learning for professional tennis match prediction and betting,” 2017.
- [2] I. McHale and A. Morton, “A bradley-terry type model for forecasting tennis match results,” *International Journal of Forecasting*, 2011.
- [3] P. Wetz, “Measuring the performance of tennis prediction models,” 2017.
- [4] R. P. Bunker and F. Thabtah, “A machine learning framework for sport result prediction,” *Applied Computing and Informatics*, 2017.