Previsão Meteorológica

Processamento de Streams
André Matos 55358
Ruben Belo 55967

Introdução

Neste projeto, o objetivo é desenvolver um modelo de previsão da meteorologia, usando o dataset escolhido previamente. Usamos várias features(precipitação, temperatura máxima, temperatura mínima e vento) para prever as condições meteorológicas(chuvisco, chuva, sol, neve e nevoeiro) num dado dia. Também usamos as mesmas colunas para prever as condições meteorológicas no dia seguinte ou nos próximos 7 dias, a quantidade de dias pode ser escolhida. O objetivo deste plano de avaliação é avaliar a desempenho e precisão do nosso modelo de previsão do tempo, usando várias métricas.

Estado da Arte

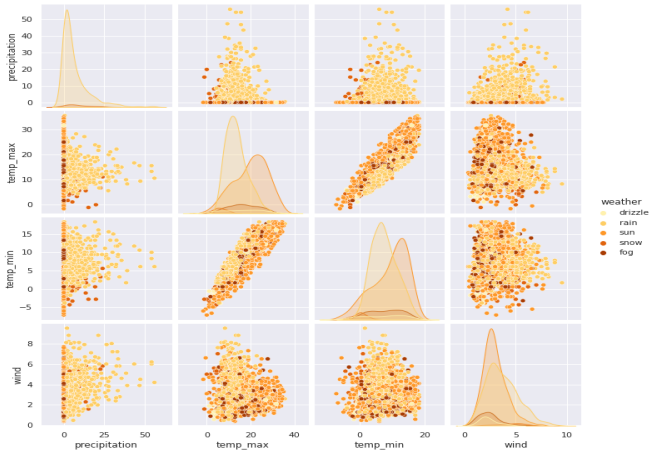
Como é sabido, há vários sitios onde se pode obter informação sobre a meteorologia, e que usam técnicas complexas de prever esta mesma, incluindo modelos que se baseiam num conjunto de equações que traduzem leis da física para descrever o comportamento hidrodinâmico da atmosfera, portanto pode-se dizer que é uma aréa com um estudo bem aprofundado.

Descrição dos Dados

O dado dataset contém as seguintes colunas: • Precipitação
• Temperatura Máxima
• Temperatura Mínima
• Vento
• Tempo

date	precipitati...	temp_max	temp_min	wind	weather
2012-01-01	0.0	12.8	5.0	4.7	drizzle
2012-01-02	10.9	10.6	2.8	4.5	rain
2012-01-03	0.8	11.7	7.2	2.3	rain
2012-01-04	20.3	12.2	5.6	4.7	rain
2012-01-05	1.3	8.9	2.8	6.1	rain
2012-01-06	2.5	4.4	2.2	2.2	rain
2012-01-07	0.0	7.2	2.8	2.3	rain
2012-01-08	0.0	10.0	2.8	2.0	sun
2012-01-09	4.3	9.4	5.0	3.4	rain
2012-01-10	1.0	6.1	0.6	3.4	rain
2012-01-11	0.0	6.1	-1.1	5.1	sun
2012-01-12	0.0	6.1	-1.7	1.9	sun

Scatter Matrix



Como esperado, vemos uma relação forte entre temperatura máxima e temperatura mínima, quanto uma aumenta, a outra também aumenta. Relativamente ás outras relações vê-se que há uma boa distribuição entre elas.

Preparação dos dados

Para a preparação dos dados foi feita primeira uma para a classificação, e depois reajustada para o problema da regressão. Primeiramente tivemos de converters os labels(condições meteorológicas) que estavam em string

para ints, e para isso usámos o LabelEncoder, depois foi só dividir o conjunto de dados em 2, um para as features e outro para as labels. Apenas isto foi necessário para a classificação. Depois para a regressão, visto que queríamos prever os dados para x dias, tivemos de dar shift das features em x linhas, uma correspondente a cada dia. Para o problema da regressão não foi necessária a coluna das labels.

Regressão Multivariada

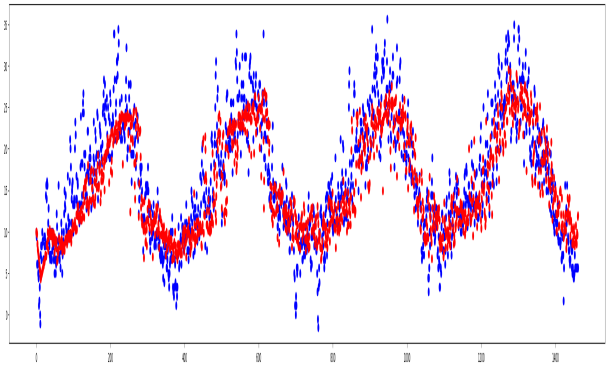
Para a Regressão Multivariada testámos todos os modelos que estavam disponíveis no riverml para regressão, desde RandomForest a modelos lineares. Rapidamente nos apercebemos que, visto que os nossos dados têm uma distribuição sinuisodal, não faria sentido usar um modelo linear, e dentro dos modelos não lineares verificámos que o KNN deu os melhores resultados. Para análise dos dados usámos tanto as métricas do riverml como uns plots feitos por nós, para melhor perceção dos resultados.
Estes foram os resultados:

Temp Max Mean Squared Error: 21.259273052424906
Temp Min Mean Squared Error: 10.439156262737457
Precipitation Mean Squared Error: 44.38395301716738
Wind Mean Squared Error: 2.1110477608099343

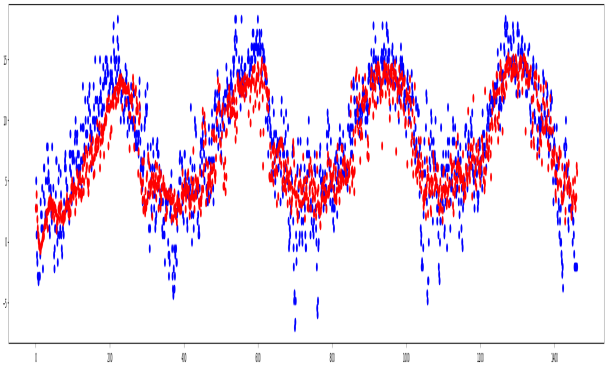
MicroAverage(MAE): 2.744524
MicroAverage(MSE): 19.408315
MicroAverage(RMSE): 4.400304
MicroAverage(R2): 0.683039

E estes foram os plots, com os pontos vermelhos a serem a previsão e os pontos azuis os reais:

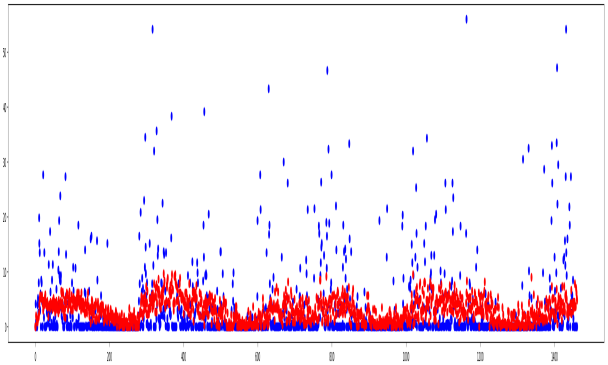
Temperatura Máxima



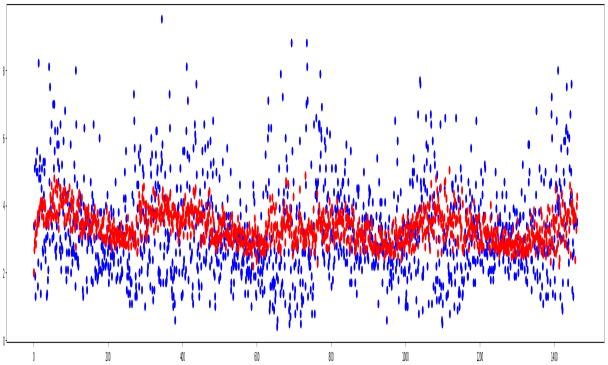
Temperatura Mínima



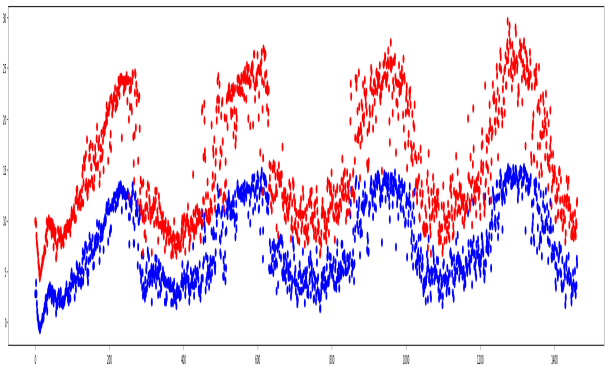
Precipitação



Vento



Temperatura Máxima Prevista vs Temperatura Mínima Prevista



Como podemos verificar nestes gráficos, os resultados previstos não estão nada longe dos resultados reais, com ênfase nas temperaturas, em que os resultados são muito idênticos. Na precipitação o modelo não se conseguiu ajustar muito aos picos de precipitação serem esporádicos, no entanto ao experimentar com outros que não fossem KNN havia uns que davam mais peso a estes picos, portanto previam melhor estes picos, no entanto no resto não tinham resultados tão bons, por isso optámos pela KNN. Relativamente às temperaturas vemos que se adaptou quase perfeitamente aos resultados, depois o vento já foi mais inconsistente, adaptou-se mas não tão bem. Com a comparação entre a temperatura máxima e a temperatura mínima vemos uma consistência.

Classificação

Na classificação, como explicado na introdução, o objetivo é prever as condições meteorológicas dados os valores de todas as outras colunas. Ao analisarmos os dados verificámos que havia class imbalance, pois maior parte das labels são chuva, sol e nevoeiro, fazendo com que o modelo ignore os chuviscos e neve. Para lidar com isto usámos todas as técnicas disponibilizadas pelo riverml, alterando também os modelos que usamos de modo a chegar ao melhor resultado em cada. De seguida mostramos os melhores resultados a que chegámos.

Class Imbalance Random Under Sampler

	Precision	Recall	F1	Support
0.0	45.45%	9.62%	15.87%	52
1.0	30.00%	11.88%	17.02%	101
2.0	66.26%	68.33%	67.28%	641
3.0	0.00%	0.00%	0.00%	26
4.0	65.19%	75.98%	70.17%	641
Macro	41.38%	33.16%	34.07%	
Micro	64.48%	64.48%	64.48%	
Weighted	61.37%	64.48%	62.05%	

64.48% accuracy

Class Imbalance Random Sampler

	Precision	Recall	F1	Support
0.0	0.00%	0.00%	0.00%	52
1.0	46.67%	6.93%	12.07%	101
2.0	66.07%	68.95%	67.48%	641
3.0	66.67%	15.38%	25.00%	26
4.0	65.06%	78.16%	71.01%	641
Macro	48.89%	33.89%	35.11%	
Micro	65.30%	65.30%	65.30%	
Weighted	61.95%	65.30%	62.04%	

65.30% accuracy

Standard

	Precision	Recall	F1	Support
0.0	50.00%	11.54%	18.75%	52
1.0	58.33%	6.93%	12.39%	101
2.0	64.61%	71.76%	68.00%	641
3.0	66.67%	23.08%	34.29%	26
4.0	66.20%	73.95%	69.86%	641
Macro	61.16%	37.45%	40.66%	
Micro	65.23%	65.23%	65.23%	
Weighted	64.39%	65.23%	62.62%	

65.23% accuracy

Ensemble ADWIN Bagging

	Precision	Recall	F1	Support
0.0	0.00%	0.00%	0.00%	52
1.0	46.67%	6.93%	12.07%	101
2.0	66.07%	68.95%	67.48%	641
3.0	66.67%	15.38%	25.00%	26
4.0	65.06%	78.16%	71.01%	641
Macro	48.89%	33.89%	35.11%	
Micro	65.30%	65.30%	65.30%	
Weighted	61.95%	65.30%	62.04%	

65.30% accuracy

Ensemble ADWIN Boosting

	Precision	Recall	F1	Support
0.0	66.67%	3.85%	7.27%	52
1.0	0.00%	0.00%	0.00%	101
2.0	64.99%	74.41%	69.38%	641
3.0	66.67%	7.69%	13.79%	26
4.0	67.13%	75.51%	71.07%	641
Macro	53.09%	32.29%	32.30%	
Micro	66.05%	66.05%	66.05%	
Weighted	61.52%	66.05%	62.13%	

66.05% accuracy

Ensemble ADA Boosting

	Precision	Recall	F1	Support
0.0	16.83%	32.69%	22.22%	52
1.0	50.00%	1.98%	3.81%	101
2.0	67.45%	58.50%	62.66%	641
3.0	21.95%	34.62%	26.87%	26
4.0	63.90%	75.66%	69.29%	641
Macro	44.03%	40.69%	36.97%	
Micro	60.78%	60.78%	60.78%	
Weighted	62.07%	60.78%	59.42%	

60.78% accuracy

Apesar da a accuracy ser maior num, não podemos apenas olhar para este valor, convém ter sempre em conta o peso que o modelo dá às classes com menos presença e também ver as outras métricas geradas. Com isto concluímos que não há um necessariamente melhor, apesar de haver uns superiores a outros.

Futuros Melhoramentos

Futuramente, uma possibilidade que se poderia fazer que seria interessante era juntar os dois. Ou seja, primeiro treina-se o modelo da classificação com os dados do dataset, depois treina-se o modelo da regressão e prevê-se os dados para, por exemplo, 7 dias. Depois para cada dia que se prevê os dados usamos o modelo da classificação para prever as condiçoes meteorológicas. Não achámos que seria relevante no âmbito de processamento de streams, no entanto era uma ideia a considerar no futuro.