Relatório sobre a atividade "11 - Prática: Predição e a Base de Aprendizado de Máquina (II)"

Lucas Gabriel Arenhardt

1. Introdução

Nesse módulo são apresentadas alguns tipos de regressão, sendo elas a linear, a polinomial e a múltipla. Também são apresentados diversos métodos utilizados para a criação de modelos em aprendizado de máquina. Todos os conceitos aprendidos nas aulas estão apresentados a seguir:

2. Regressão

2.1. Regressão Linear

É uma técnica utilizada para modelar a relação entre uma variável dependente e uma independente. O objetivo é encontrar a melhor linha reta que descreva essa relação, permitindo predizer valores novos com base nisso. Para medir o quão bem a reta é capaz de predizer os pontos, usa-se o coeficiente de determinação (R^2), onde o valor 0 indica uma péssima correlação e o valor 1 uma ótima correlação.

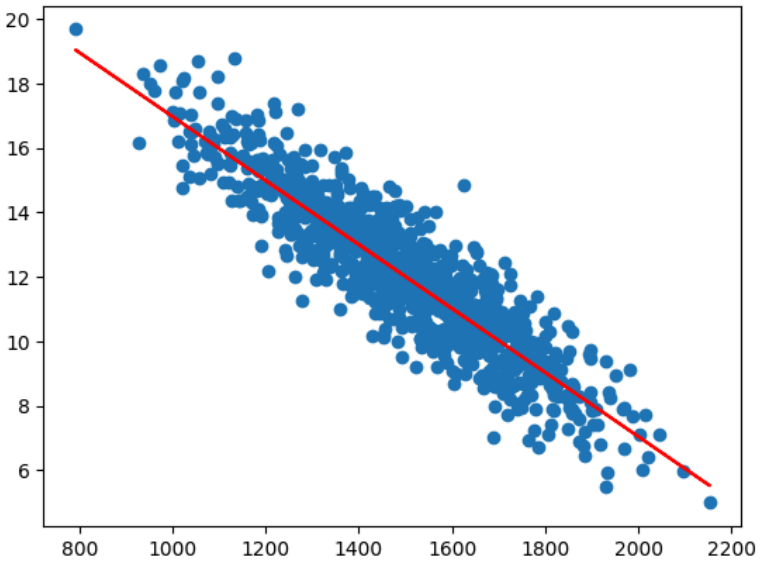

Imagem 1 - A linha vermelha representa a reta de regressão.

2.2. Regressão Polinomial

Essa forma de regressão é utilizada quando uma reta não se encaixa corretamente na relação, sendo assim necessária uma curva. Quanto maior o grau do polinômio utilizado, mais complexa será a curva.

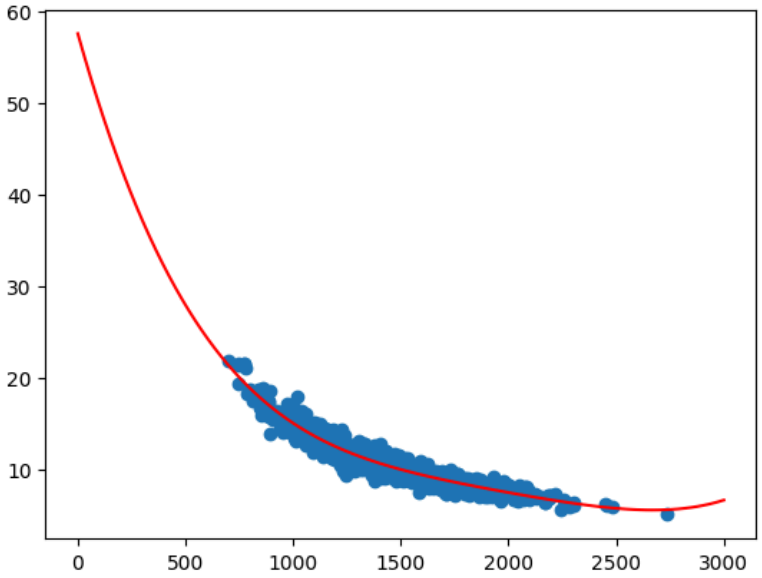

Imagem 2 - A linha vermelha representa o polinômio de regressão.

2.3. Regressão Múltipla

A regressão múltipla permite analisar a relação entre uma variável dependente e múltiplas variáveis independentes, sendo uma extensão das demais regressões.

3. Multi-Level Models

Os Multi-Level Models são utilizados para analisar dados que possuem uma estrutura hierárquica ou aninhada, ou seja, dados que possuem influência de outros dados.

4. Aprendizado de Máquina

4.1. Aprendizado Não Supervisionado

No aprendizado não supervisionado, o modelo é treinado com um conjunto de dados não rotulados. Aqui, o objetivo é descobrir padrões sem nenhuma orientação explícita. É muito útil para quando você não sabe exatamente o que está buscando.

4.2. Aprendizado Supervisionado

No aprendizado supervisionado, o modelo é treinado com um conjunto de dados rotulados. Ou seja, o algoritmo aprende com base em respostas “corretas” dadas pelo programador.

5. Naive Bayes

O Naive Bayes é uma família de algoritmos de classificação baseados no Teorema de Bayes (já abordado anteriormente durante o curso).

6. K-Means Clustering

É um algoritmo de aprendizado não supervisionado utilizado para particionar um conjunto de dados em K grupos (clusters), onde cada dado pertence ao cluster com o centróide mais próximo.

7. Entropia

A entropia mede o quão diferente é um Dataset. O valor 0 quer dizer que todas as classes do Dataset são iguais. A entropia é alta se as classes são diferentes.

8. Árvores de Decisão

É uma ferramenta de aprendizado de máquina que funciona dividindo repetidamente os dados em subconjuntos com base em uma série de condições, representadas por uma estrutura de árvore. Assim, cada nó representa uma decisão e cada ramo a sua respectiva consequência.

9. Ensemble Learning

É uma técnica de aprendizado de máquina onde múltiplos modelos são treinados para resolver o mesmo problema. A ideia é que, combinando vários modelos, é possível obter resultados superiores do que qualquer um dos modelos individuais. Algumas abordagens de Ensemble Learning são:

9.1. Bagging

Vários modelos são treinados usando diferentes subconjuntos de dados de treinamento. A predição final é obtida pela votação dos modelos individuais.

9.2. Boosting

Os modelos são treinados sequencialmente, de modo que o próximo tende a corrigir os erros do anterior.

9.3. Bucket of Models

Diversos tipos de modelos são treinados e ao final é escolhido aquele que melhor consegue lidar com os dados.

9.4. Stacking

Diversos tipos de modelos são treinados. Ao final, as predições desses modelos são utilizadas como entradas para um outro modelo, combinando todos.

10. XGBoost

É uma biblioteca de aprendizado de máquina otimizada, projetada para ser altamente eficiente. É uma implementação de gradient boosting, uma técnica que combina múltiplos modelos de aprendizado de máquina (árvores) para formar um modelo ainda melhor.

O XGBoost possui várias funções interessantes, como essas: é capaz de lidar automaticamente com valores ausentes e utiliza processamento paralelo.

Para utilizá-lo é necessário configurar seus hiperparâmetros (booster, objective, eta, max_depth, min_child_weight, entre outros).

11. Support Vector Machines (SVM)

O objetivo principal do SVM é encontrar o hiperplano que melhor separa os dados em diferentes classes. Esse método funciona muito bem para quando há várias dimensões a serem trabalhadas.

12. Conclusão

Todos os conceitos apresentados nas aulas são de grande importância para o aprendizado de máquina. Eu já havia trabalhado anteriormente com regressão, porém sua aplicação em python foi algo novo para mim. As práticas de criação de modelos foram essenciais para o entendimento do conteúdo.