Föreläsning 05

Övningar

Predicting diamond prices. Använd diamon.csv för att prediktera target *price*. Det är tre variabler som är kategoriska, cut, color och calrity. Skapa dummy-variabler av dessa. Kommer visas i lektionen. Följ de sex stegen för att prediktera som vi har gått igenom och prediktera med modellerna NULL, MLR, KNN och LASSO. Hur bra predikterar modellerna? Jämför resultaten visuellt med en valbar error metric.

Frågor

1. Vad är intentionen bakom K-Nearest Neighbor?
2. Vilka steg finns till K-Nearest Neighbor?
3. Vad står ”K” för?
4. Vika viktningsfunktioner finns till ”K”?
5. Vilken populär metrisk distans använder man?
6. Vilka antaganden gör man vid KNN?
7. Vad är viktigt att tänka på när man använder KNN?
8. Hur fungerar Lasso regression?
9. Hur skiljer den sig från Multiple Linear regression?
10. Vilka egenskaper har Lasso regression?
11. Hur fungerar Alpha i Lasso regression?
12. Vad betyder Coss-Validation?
13. Vad menas med Overfitting?
14. Vad gör att risken för Overfitting ökar?
15. Vad är Regularization?

Länkar

- KNN https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761
- Euclidean, Minkowski distance https://www.youtube.com/watch?v=ofIfXMPem2M&ab_channel=LearningMonkey
- Overfitting - https://blog.minitab.com/blog/understanding-statistics/how-to-avoid-overfitting-your-regression-model
- Overfitting - https://statisticsbyjim.com/regression/overfitting-regression-models/
- Regularization - https://towardsdatascience.com/regularization-an-important-concept-in-machine-learning-5891628907ea
- Lasso regression https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b