Inlämningsuppgift Prediktiv analys

Mars 2022

Generell information

Inlämning sker i form utav en Jupyter Notebook-fil (.ipynb) på PingPong. Deadline 2022-03-18 kl 23:55. Filen namnges med namn, Prediktiv analys och betygsönske, tex *Eva_Hegnar_Prediktiv_analys_VG.ipynb*.

Kom ihåg att kopiera kod rakt av räknas som fusk, men man får hämta inspiration. **Ange källor!** Inlämningen är **individuell**.

Betygskriterier

G

- Kunna på ett grundläggande sätt förklara regression och klassificering
- Kunna på ett grundläggande sätt tillämpa prediktiva analysalgoritmer och modeller
- Kunna på ett grundläggande sätt tillämpa Machine learning metoder
- Kunna på ett grundläggande sätt skapa visualisering av prediktiva analyser
- Kunna på ett grundläggande sätt utföra prediktiv analys från dataset

VG

- Uppnått kraven för betyget Godkänd.
- Kunna på ett fördjupat sätt tillämpa prediktiva analysalgoritmer och modeller
- Kunna på ett detaljerat sätt tillämpa Machine learning metoder
- Kunna på ett kreativt sätt skapa visualisering av prediktiva analyser
- Kunna på ett självständigt sätt utföra prediktiv analys från dataset

Introduktion

I denna inlämningsuppgift ska ni jobba med regression. Ni ska använda prediktiv_data.csv för att prediktera kolumneen target. Ni väljer själv vilka features ni inkluderar i modellerna och vilka modeller ni ska använda. Ni ska prediktera i Python genom att följa de stegen vi har gått igenom för att bygga en prediktiv modell och analysera resultaten:

1. Förbereda datan
2. Importera modellen
3. Skapa instans av modellen
4. Träna modellen med träningsdatan
5. Utvärdera modellen
6. Gör predikteringar

Att prediktera med data är en iterativ process, det kan alltså hända ni får hoppa fram och tillbaka mellan stegen flera gånger innan resultaten blir bra. Vilka val ni har gjort och varför kommenterar ni i notebooken. Använd Markdown för detta!

G

I tillägg till att följa stegen ovanför ska åtminstone vara med för godkänt (ni får alltså ta med mer om det önskas):

- Utforska datan – Använd det ni har lärt tidigare om exploratory data analysis med att undersöka data med scatter plots och statistisk om datan (särskilt om target).
- Förberedelse av data – Saknas det värden? Är några av variablerna kategoriska så det måste skapas dummy variabler?
- Välj en machine learning modell att prediktera med.
- Utvärdering av modellen – vad är resultatet av träningen och testningen av modellen? Välj en error metrics för att utvärdera modellen

VG

För VG ska ni ha uppnått kraven för betyget Godkänd samt:

- Välj tre olika modeller och jämför resultaten. Skapa också en NULL modell som predikterar genom att beräkna genomsnittet som visad i lektionen.
- Det är olika antagen som gäller för de olika modellerna. Uppfylls dessa antagen? Undersöka datan noggrant. Kan detta motivera hur väl eller dåligt modellerna fungerar på att prediktera? Ni får göra feature engineering, dvs ändra i datan, för att uppnå bättre resultat i predikteringen.
- Gör val av input features till modellen och motivera varför. Här ska ni lägga lägga mer krut på exploratory data analysis delen. Finns det till exempel outliers som behöver hanteras? Är variablerna korrelerade? Gör feature selection och se om ni kan uppnå ett bättre resultat.
- Resultat av prediktions analysen. Ta fram flera error metrics, plotta resultat om det passar för modellen. Kritiskt analysera resultatet av modelleringarna.
- Använd Markdown i notebooken för alla analyser och reflektioner.