Deep Learning

2023-11-24

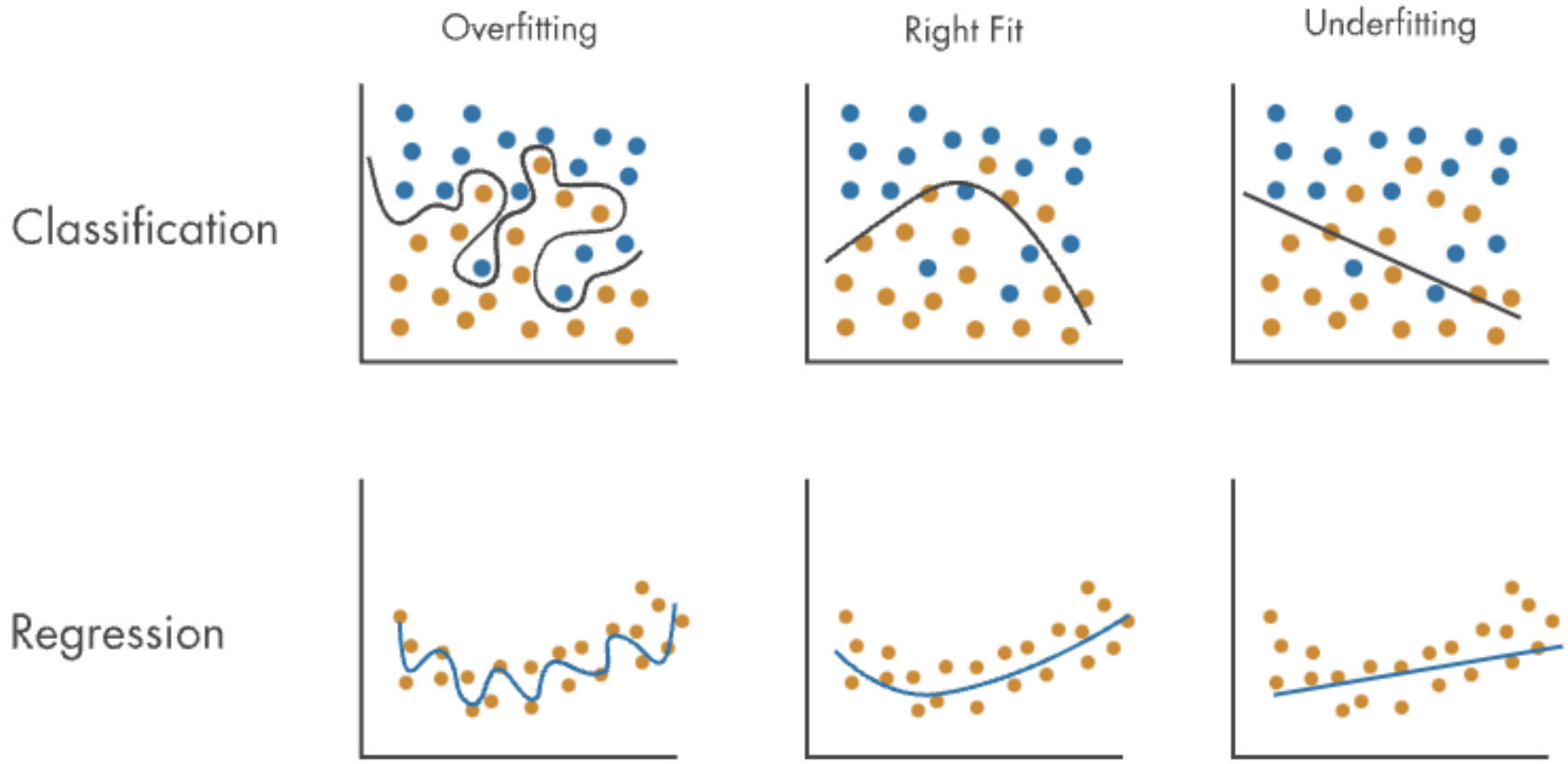
Ämne: Bias/Variance, overfit/underfit, early stopping

Smartr

Agenda

- Underfit och overfit
- Bias vs Variance tradeoff
- Regularisering – early stopping

Underfitting vs Overfitting



Dataset terminologi

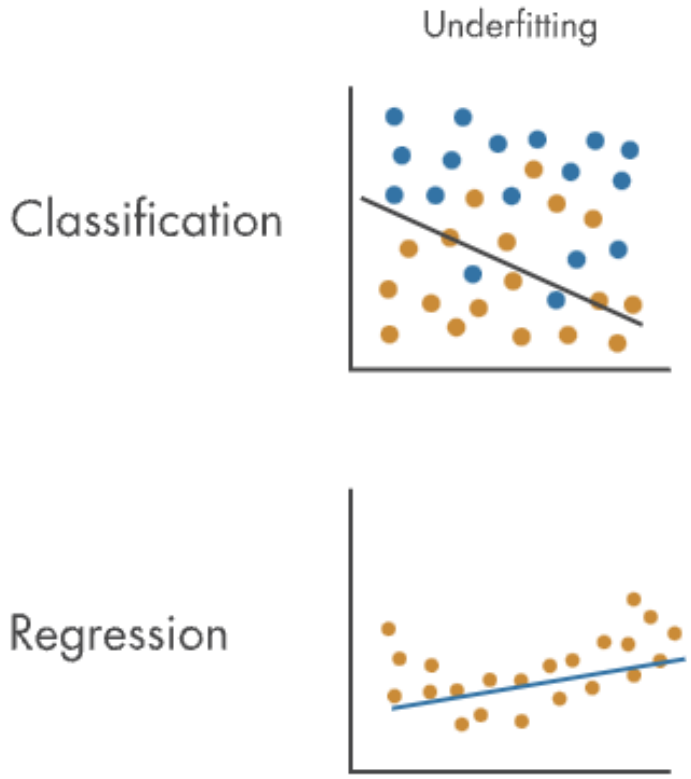
- Det råder inte helt 100% koncensus kring namn av dataset så låt oss reda ut vad vi använder i denna kursen. (Om inget annat anges.)

Dataset	Mängd av total data	Används för
Träningsdata	Ca 70%	Träna nätverket
Valideringsdata	Ca 20%	Hyperparameteroptimering.
Testdata	Ca 10%	Slutlig jämförelse av modellen mot andra modeller. Du får inte röra testdatan förrän modellen är helt klar.

Underfitting

- Det innebär att modellen ej fångat mönstren i datan och inte lärt sig.

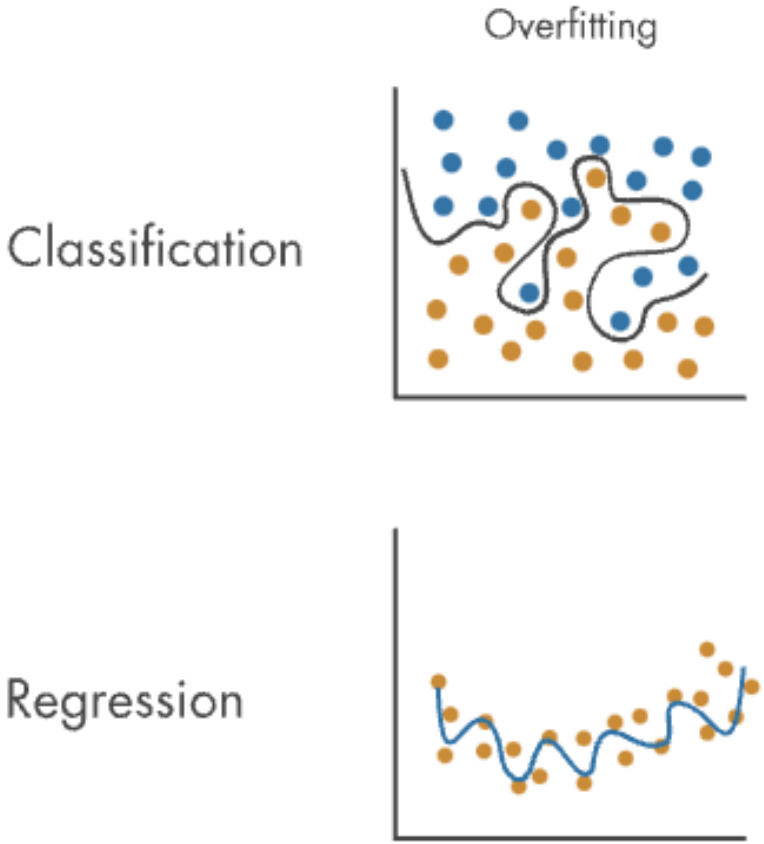
- Det kan bero på:
  - Modellen är för simpel för att fånga sambandet.
  - Modellen har inte konvergerat (av någon anledning...).
  - Modellen regulariseras för hårt.* (återkommer till detta)



Overfitting

- Modellen har lärt sig datan utantill och generaliserar inte väl till osedd data.

- Det kan bero på
  - Modellen är för komplex och kan memorera datan.
  - Datamängden är för liten eller innehåller för lite relevant information.



Underfitting vs Overfitting

Typ av error	Overfitting	Right fit	Underfit
Training error	låg	låg	hög
Test error	hög	låg	hög

Hur undviker man overfitting?

- Reducera modellkomplexiteten
  - Mer regularisering (mer nästa vecka)
    - T.ex. early stopping!
  - Förenkla arkitekturen

- Förbättra datakvalitéten och öka mängden data
  - Data augmentering
  - Data generering
  - Datastädning
  - Reducera antalet data features (input dimensioner).

Hur undviker man underfitting?

- Öka modellkomplexiteten
  - Minska regularisering
  - Utöka arkitekturen

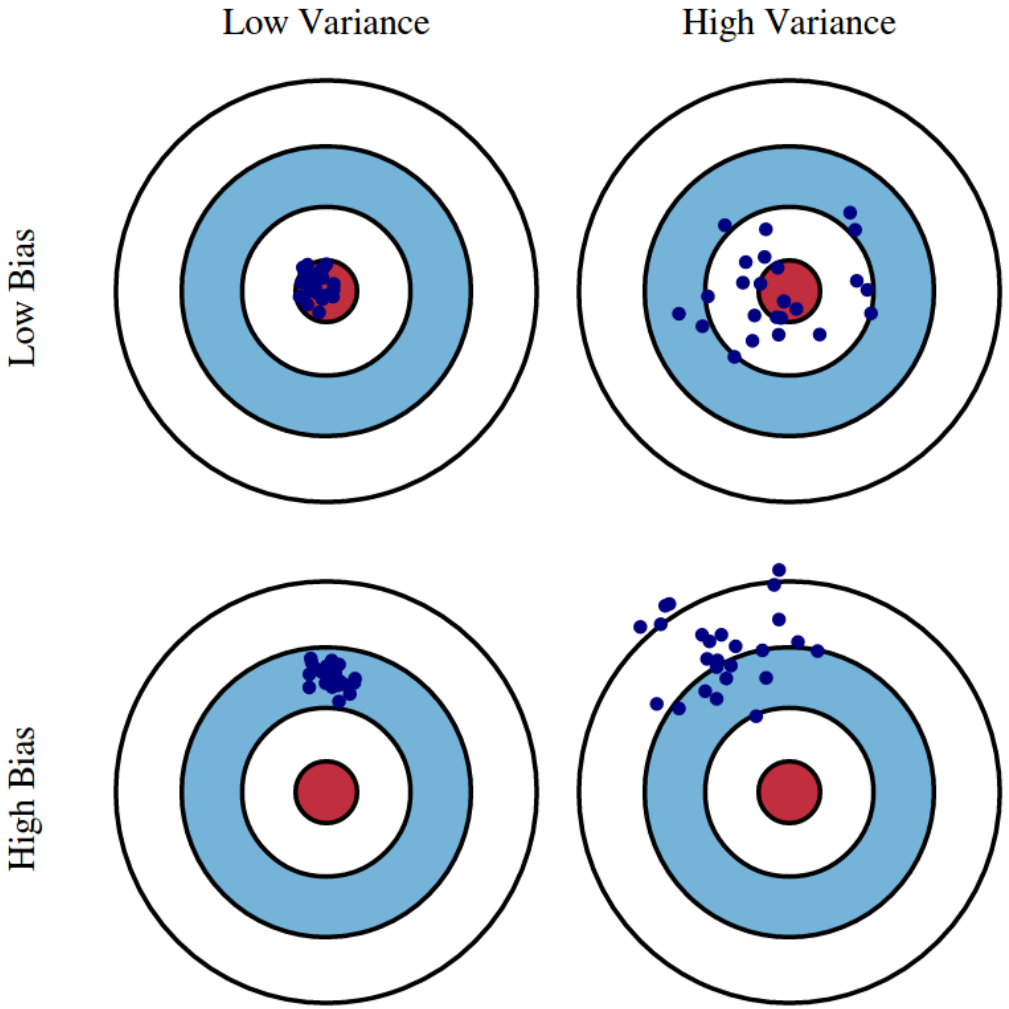
- Förbättra datakvalitéten och öka mängden data
  - Data augmentering
  - Data generering
  - Datastädning
    - Rensa bort ”noise” i datan.
  - Öka antalet data features (input dimensioner).

Bias and Variance tradeoff

- Klassiskt problem i supervised learning

- Hög bias innbär att modellen ej lärt sig datan och *underfitat*
- Hög varians innebär ofta att modellen har *overfitat.*

- Vi vill hitta en modell som både har låg bias och låg varians.
- Det innebär att modellen lärt sig från datan men också generaliserar bra till osedd data.



r

Regulariseringstekniker

- Early stopping är den enklaste och absolut vanligaste regulariseringstekniken.

- Andra regulariseringstekniker går vi igenom i nästa vecka.

Early stopping

- Intuition: Vi avbryter inlärningen innan modellen lärt sig datasetet utantill. Vi hoppas att den då generaliserar väl.

- Det sker när error för valideringsdatan, som vi inte tränar på och därför är osedd data, slutar gå ner.