STOCKHOLMS UNIVERSITET
MATEMATISKA INSTITUTIONEN
Avd. Beräkningsmatematik

TENTAMEN
Introduktion till maskininlärning
17 mars 2021

Tentamen för kursen
Introduktion till maskininlärning
17 mars 2021 8–14

Ansvarig lärare: Sebastian Rosengren, rosengren@math.su.se
Examinator: Lars Arvestad
Hjälpmedel: Kurslitteratur och föreläsningsanteckningar.
Återlämning: Information om återlämning skickas ut via kursforum.
Lösningar: Finns på kursens hemsida efter skrivtidens slut.

Varje korrekt löst uppgift ger 12 poäng. E: 30p, D: 35p, C: 40p, B: 45p, A: 50p.

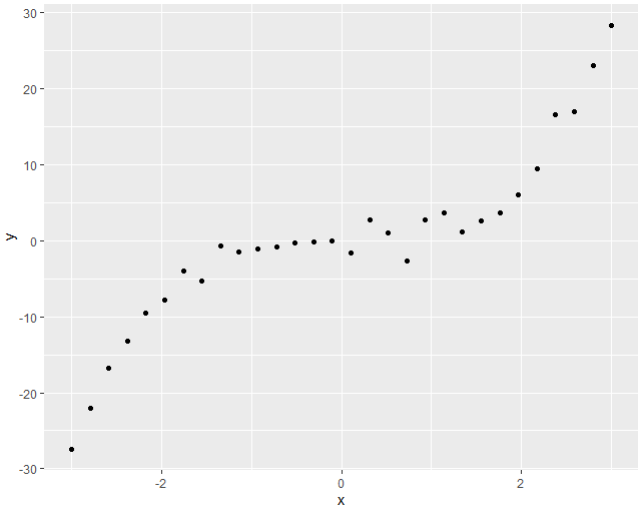
Svar ska motiveras och resonemang skall vara klara och tydliga att följa.

Vidare är det en god idé att läsa bifogad fil:
`Praktiskt information angående inlämning av tentamen`.

Lycka Till!

Uppgift 1

Antag att vi har följande 30 datapunkter.



Antag vidare att $y_i = f_\theta(x_i) + \epsilon_i$, där $\epsilon_i,\ i = 1, 2, \ldots, 30$ är oberoende slumpfel. Ange för nedanstående modeller om de har hög eller låg *bias*, samt hög eller låg *varians*. Endast svar räcker.

a.) $f_\theta(x) = \alpha + \sum_{i=1}^{30} \beta_i x^i$ (2 p)

b.) $f_\theta(x) = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$ (2 p)

c.) $f_\theta(x)$ är ett neuralt nät med ReLu aktivering, fem gömda lager, med 10 noder per lager. (2 p)

d.) $f_\theta(x) = \frac{1}{30} \sum_{x_i \in N_{30}(x)} y_i$, alltså en knn regression med $k = 30$. (2 p)

e.) $f_\theta(x) = \sum_{x_i \in N_1(x)} y_i$, alltså en knn regression med $k = 1$. (2 p)

f.) $f_\theta(x) = c_1 1\{x \in R_1\} + c_2 1\{x \in R_2\}$, alltså ett beslutsträd med två regioner. (2 p)

Uppgift 2

a.) Antag att du ska träna en modell som ska användas för prediktion. Definiera *mean absolute error (MAE)* och *mean absolute percentage error (MAPE)*, samt ge ett exempel på en situation då MAE är bättre att använda än MAPE. (2 p)

b.) Vad är det *primära* syftet med bagging, och varför är beslutsträd lämpliga för algoritmen? (2 p)

c.) Varför är det ofta viktigt att standardisera eller normalisera data i k-nearest neighbors modeller? (2 p)

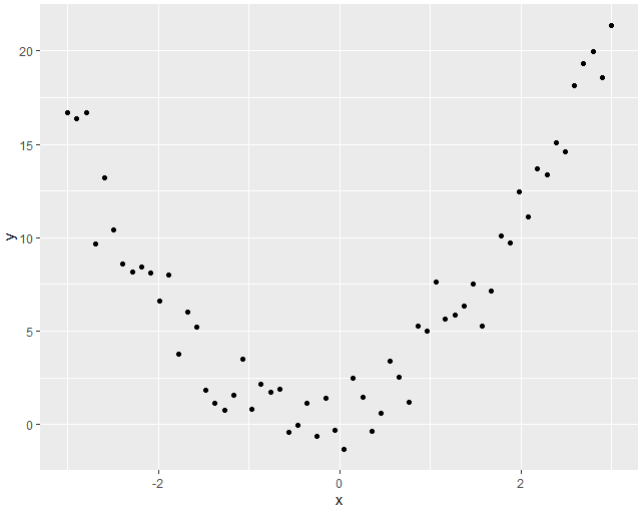
d.) Antag att $\mathbb{P}(y = 1; x)$ har ett icke-monotont beroende på x. Definiera modellen *enkel logistisk regression*, och förklara varför den inte är lämplig för denna situation. (2 p)

e.) Definiera vad som menas med en optimal policy π inom Reinforcement Learning. (2 p)

f.) Beskriv *value iteration* och förklara hur det kan användas för att konstruera en optimal policy π^*. (2 p)

Uppgift 3

Antag att vi har följande data.



Data är också bifogad i filen `data_uppgift3.csv`.

Låt $f_\theta(x) = \theta^T \phi(x)$, där $\phi : \mathbb{R} \to \mathbb{R}^k$ är en transformation av data. Vi ansätter en linjär regressionsmodell, $y_i = f_\theta(x_i) + \epsilon_i$, där $\epsilon_i,\ i = 1, 2, \ldots$ är oberoende slumpfel.

a.) Ange en lämplig transformation $\phi(x)$ av data. (3 p)

b.) Beräkna parameterskattningarna för modellen $f_\theta(x_i) = \alpha + \beta_1 x$, samt beräkna *mean squared error* för modellen på träningsdata. (3 p)

c.) För en godtycklig linjär regressionsmodell skissa algoritmen för k-fold korsvalidering, samt förklara syftet med korsvalidering. (4 p)

d.) Antag att du anpassat en linjär regressionsmodell som verkar ha overfittat data. Ge exempel på två metoder som kan minska detta problem, samt ge en översiktlig förklaring av metoderna. (2 p)

Uppgift 4

Låt nu $f_\theta(x) \in \mathbb{R}$ vara ett specifikt neuralt nätverk för regression (endimensionell output) med ReLu aktivering. Antag att $x \in \mathbb{R}^5$ (femdimensionell input), samt att $f_\theta(x)$ har 3 gömda lager, med 12, 13, respektive 14 noder.

För **a.)** och **b.)** antag att modellen ska tränas och användas för prediktion.

a.) Ge ett typiskt exempel på en situation då modellen förmodligen är lämplig. (2 p)

b.) Ge ett typiskt exempel på en situation då modellen förmodligen är olämplig. (2 p)

c.) Beräkna hur många parametrar det finns i modellen. (4 p)

d.) Antag att modellparametrarna är kända. Förklara hur värdena på de gömda lagerna h^1, h^2, h^3, och outputen $f_\theta(x)$, beräknas givet en input x. (4 p)

Uppgift 5

a.) Beskriv detaljerat, steg för steg, algoritmen för k-nearest neighbor klustring. Använd både text och matematisk notation. (3 p)

b.) Gör din egen implementation av algoritmen i valfritt programmeringsspråk genom att skapa en funktion `k_means(X, centers)`. Funktionen ska ta som inputs ett dataset $X \in \mathbb{R}^d$ *och en matris med initiala värden på mittpunkterna för klusterna*, notera att *centers* är en $k \times d$-matris. Du får självklart använda redan färdiga funktioner för att räkna ut medelvärden osv. Koden ska bifogas, och vara dokumenterad. (3 p)

c.) Använd algoritmen du skapade i **b.)** för att klustra datasetet i filen `data_uppgift5.csv` i två kluster. Om du inte löste **a.)**, får du använda en färdig implementation `kmeans`. Redovisa resultatet genom att plotta datapunkterna och färga dem givet klusterindelningen. (3 p)

d.) Analysera resultatet i **c.)**. Ser klustringen rimlig ut? Förklara resultatet. Ge exempel på en algoritm för klustring som skulle kunna vara mer lämplig för data av denna karaktär. Motivera! (3 p)