MAT1080 H25
Oscar Brovold

Løsningsforslag kapittel 4

Oppgaver relatert til maskinlæring

Oppgave 34

Vi ser her på den lineære funksjonen:

$$f(x_1, x_2, ..., x_D) = \sum_{d=1}^{D} w_d x_d + b.$$

a)

Vi observerer at argumentet til funksjonen kan skrives som en vektor $\boldsymbol{x} = [x_1, x_2, \ldots, x_D]^\top$, slik at vi får

$$f(\boldsymbol{x}) = \sum_{d=1}^{D} w_d x_d + b.$$

I tillegg utnytter vi at koeffisientene til den lineære funksjonen også kan samles i en vektor $\boldsymbol{w} = [w_1, w_2, \ldots, w_D]^\top$. En nyttig innsikt, som også løser oppgaven, er følgende sammenheng:

$$\boldsymbol{w}^\top \boldsymbol{x} \;=\; \begin{bmatrix} w_1 & w_2 & \cdots & w_D \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix} \;=\; w_1 x_1 + w_2 x_2 + \cdots + w_D x_D \;=\; \sum_{d=1}^{D} w_d x_d.$$

Altså har vi nå funksjonen med vektornotasjon:

$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b.$$

b)

Oppgaven ber oss forklare at mengden $\left\{ \boldsymbol{x} \in \mathbb{R}^D \;\middle|\; f(\boldsymbol{x}) = b \right\}$ står ortogonalt på $\boldsymbol{w}$. Mengden består av alle vektorer $\boldsymbol{x} \in \mathbb{R}^D$ som gjør at funksjonsverdien er konstanten b. Vi oppnår $f(\boldsymbol{x}) = b$ dersom $\boldsymbol{w}^\top \boldsymbol{x} = 0$, altså når indreproduktet (prikkproduktet) er lik 0. At $\boldsymbol{w}^\top \boldsymbol{x} = 0$ er det samme som at vektoren $\boldsymbol{x}$ står normalt på $\boldsymbol{w}$.

Oppgave 35

I maskinlæring er vi ofte interessert i representasjoner. Altså gitt data (f.eks. tekst) er vi interessert i følgende spørsmål; kan vi oppsummere/representere innholdet i teksten på en fornuftig måte. En naiv fremgangsmåte kunne vært å oversette hvert ord i teksten med et tall og samle disse i en vektor, dette ville da vært en vektorrepresentasjon av teksten. En smartere fremgangsmåte ville vært å definere en funksjon, $f(\cdot)$, hvor inndata til funksjonen er teksten og resultatet er en lært representasjon, som forhåpentligvis representerer det semantiske innholdet på en mer fornuftig måte enn den naive fremgangsmåten. Funksjonen $f(\cdot)$ er her ofte en form for et nevralt nettverk.

Cosinuslikhet gir oss en meningsfull måte å sammenlikne to slike vektorrepresentasjoner. Anta at du har to tekster, tekst 1 og tekst 2 og en maskinlæringsmodell $f(\cdot)$, du kjører tekst 1 og tekst 2 gjennom f og får ut to representasjoner $\boldsymbol{v}_1$ og $\boldsymbol{v}_2$, altså:

$$f(\text{tekst 1}) = \boldsymbol{v}_1 \qquad \text{og} \qquad f(\text{tekst 2}) = \boldsymbol{v}_2.$$

Vi har at $\cos(\boldsymbol{v}_1, \boldsymbol{v}_2) \in [-1, 1]$ og at to vektorer er like dersom $\cos(\boldsymbol{v}_1, \boldsymbol{v}_2)$ er nærme 1 og ulike dersom de er nærme -1. Forutsatt at f produserer meningsfulle representasjoner gir derfor cosinuslikhet en meningsfull måte å sammenlikne tekster på.

Vi starter med å se på noen egenskaper ved $\cos(\boldsymbol{v}_1, \boldsymbol{v}_2)$:

a)

Vi antar $\boldsymbol{v}_1, \boldsymbol{v}_2 \neq \mathbf{0}$. Definisjonen av cosinus mellom to vektorer er

$$\cos(\boldsymbol{v}_1, \boldsymbol{v}_2) = \frac{\boldsymbol{v}_1^\top \boldsymbol{v}_2}{\|\boldsymbol{v}_1\|_2 \, \|\boldsymbol{v}_2\|_2}.$$

Vi viser nå at dette er lik indreproduktet av de normaliserte vektorene:

$$\left(\frac{\boldsymbol{v}_1}{\|\boldsymbol{v}_1\|_2}\right)^\top \left(\frac{\boldsymbol{v}_2}{\|\boldsymbol{v}_2\|_2}\right) = \frac{1}{\|\boldsymbol{v}_1\|_2} \, \boldsymbol{v}_1^\top \, \frac{1}{\|\boldsymbol{v}_2\|_2} \, \boldsymbol{v}_2 = \frac{1}{\|\boldsymbol{v}_1\|_2 \, \|\boldsymbol{v}_2\|_2} \, \boldsymbol{v}_1^\top \boldsymbol{v}_2.$$

Dermed

$$\cos(\boldsymbol{v}_1, \boldsymbol{v}_2) = \frac{\boldsymbol{v}_1^\top \boldsymbol{v}_2}{\|\boldsymbol{v}_1\|_2 \, \|\boldsymbol{v}_2\|_2} = \left(\frac{\boldsymbol{v}_1}{\|\boldsymbol{v}_1\|_2}\right)^\top \left(\frac{\boldsymbol{v}_2}{\|\boldsymbol{v}_2\|_2}\right).$$

Merk: Hvis én av vektorene er nullvektoren, er uttrykket ikke definert siden vi ikke kan dele på 0.

b)

Vi tar nå for oss vektorene:

$$\boldsymbol{w}_1 = [1, 0, 3]^\top \qquad \text{og} \qquad \boldsymbol{w}_2 = [2, -1, 1]^\top.$$

Vi skal finne cosinuslikhet ved å normalisere først. Vi har:

$$\hat{\boldsymbol{w}}_1 = \frac{\boldsymbol{w}_1}{\|\boldsymbol{w}_1\|_2} = \frac{1}{\sqrt{1^2 + 0^2 + 3^2}} \begin{bmatrix} 1 \\ 0 \\ 3 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{10}} \\ 0 \\ \frac{3}{\sqrt{10}} \end{bmatrix},$$

og

$$\hat{\boldsymbol{w}}_2 = \frac{\boldsymbol{w}_2}{\|\boldsymbol{w}_2\|_2} = \frac{1}{\sqrt{2^2 + (-1)^2 + 1^2}} \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{2}{\sqrt{6}} \\ -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{bmatrix}.$$

Vi får da:

$$\begin{aligned}
\cos(\boldsymbol{w}_1, \boldsymbol{w}_2) &= \hat{\boldsymbol{w}}_1^\top \hat{\boldsymbol{w}}_2 \\
&= \begin{bmatrix} \frac{1}{\sqrt{10}} & 0 & \frac{3}{\sqrt{10}} \end{bmatrix} \begin{bmatrix} \frac{2}{\sqrt{6}} \\ -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{bmatrix} \\
&= \frac{1}{\sqrt{10}} \frac{2}{\sqrt{6}} + 0 \cdot \Big(-\frac{1}{\sqrt{6}} \Big) + \frac{3}{\sqrt{10}} \frac{1}{\sqrt{6}} \\
&= \frac{5}{\sqrt{60}} = \frac{5}{2\sqrt{15}} \approx 0.6455.
\end{aligned}$$

c)

Denne oppgaven tar for seg en situasjon hvor du har 100 tekster, og er interessert i å finne ut hvilke tekster som er like og hvilke som er ulike. Du har trent en maskinlæringsmodell, $f(\cdot)$, og har produsert representasjoner av tekstene ved hjelp av $f(\cdot)$. Vi skal nå se på hvordan vi kan utnytte matrisemultiplikasjon for å beregne cosinuslikheten mellom alle representasjonene.

Spesifikt for oppgaven antar vi at hver tekst har en representasjon, som er en vektor $\boldsymbol{v}_i \in \mathbb{R}^{64}$, altså $\boldsymbol{v}_i = [v_{i1}, v_{i2}, \ldots, v_{i64}]^\top$. Siden vi har 100 tekster har vi $i = 1, 2, \ldots, 100$. Vi definerer nå en matrise V på følgende måte:

$$V = \begin{bmatrix} \boldsymbol{v}_1^\top \\ \boldsymbol{v}_2^\top \\ \vdots \\ \boldsymbol{v}_{100}^\top \end{bmatrix} = \begin{bmatrix} v_{1,1} & v_{1,2} & \cdots & v_{1,64} \\ v_{2,1} & v_{2,2} & \cdots & v_{2,64} \\ \vdots & \vdots & \ddots & \vdots \\ v_{100,1} & v_{100,2} & \cdots & v_{100,64} \end{bmatrix} \in \mathbb{R}^{100\times 64}.$$

d)

Vi ser nå på matrisemultiplikasjonen $VV^\top$, der

$$V = \begin{bmatrix} \boldsymbol{v}_1^\top \\ \boldsymbol{v}_2^\top \\ \vdots \\ \boldsymbol{v}_{100}^\top \end{bmatrix}.$$

Da får vi:

$$VV^\top = \begin{bmatrix} \boldsymbol{v}_1^\top \\ \boldsymbol{v}_2^\top \\ \vdots \\ \boldsymbol{v}_{100}^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \cdots & \boldsymbol{v}_{100} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1^\top \boldsymbol{v}_1 & \boldsymbol{v}_1^\top \boldsymbol{v}_2 & \cdots & \boldsymbol{v}_1^\top \boldsymbol{v}_{100} \\ \boldsymbol{v}_2^\top \boldsymbol{v}_1 & \boldsymbol{v}_2^\top \boldsymbol{v}_2 & \cdots & \boldsymbol{v}_2^\top \boldsymbol{v}_{100} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{v}_{100}^\top \boldsymbol{v}_1 & \boldsymbol{v}_{100}^\top \boldsymbol{v}_2 & \cdots & \boldsymbol{v}_{100}^\top \boldsymbol{v}_{100} \end{bmatrix}.$$

Elementet på rad i og kolonne j i produktet $VV^\top$ er gitt ved:

$$(VV^\top)_{ij} = \boldsymbol{v}_i^\top \boldsymbol{v}_j.$$

Dette følger direkte fra definisjonen av matriseproduktet, der hvert element i resultatet fås ved indreproduktet (prikkproduktet) mellom rad i i V og kolonne j i $V^\top$. Dermed inneholder $VV^\top$ alle parvise indreprodukter mellom tekstrepresentasjonene.

Hvis vi nå normaliserer hver rad i V først, og kaller denne matrisen for $\hat{V}$ får vi:

$$(\hat{V}\hat{V}^\top)_{ij} \;=\; \hat{\boldsymbol{v}}_i^\top \hat{\boldsymbol{v}}_j \;=\; \cos(\boldsymbol{v}_i, \boldsymbol{v}_j).$$

Altså har vi nå en matrise $\hat{V}\hat{V}^\top$ hvor hver indeks (i, j) inneholder et likhetsmål mellom tekst i og tekst j.

e)

Vi skal i denne oppgaven se på summen $\sum_{i=1}^{100} (\hat{V}\hat{V}^\top)_{ii}$.

Altså skal vi summere opp diagonalelementene i matrisen $\hat{V}\hat{V}^\top$. Siden diagonalelementene er cosinuslikhet mellom to like vektorer, som er 1, har vi:

$$\sum_{i=1}^{100} (\hat{V}\hat{V}^\top)_{ii} = \sum_{i=1}^{100} 1 = 100.$$

f)

En informatiker er ofte interessert i hvor effektiv en beregning/algoritme er. Vi ser derfor i denne oppgaven på hvor mange indreprodukter denne matrisemultiplikasjonen $\hat{V}\hat{V}^\top$ krever.

I tilfellet over hvor vi har 100 tekster krever vi $100 \cdot 100 = 100^2 = 10,000$ indreprodukter. For n tekster krever vi n^2 indreprodukter. Dette skalerer seg svært dårlig. Dersom vi øker til bare $10,000$ tekster trenger vi $10,000 \cdot 10,000 = 10^8 = 100,000,000$ indreprodukter.

Bemerkning: Cosinuslikhetsmatrisen $S = \hat{V}\hat{V}^\top$ er symmetrisk, altså $S_{ij} = S_{ji}$. Dermed finnes det bare $\frac{n(n+1)}{2}$ *distinkte* elementer (inkludert diagonalen). Dette er noe vi kan utnytte i en faktisk implementasjon.