Machine learning - DTU

Rapport

Anna Louise Hansen

Fra Udviklings- og Forenklingsstyrrelsen

Chapter 1

Part I

1.1 Beskrivelse af data

Alle boligejere i Danmark betaler en skat, ejendomsværdiskat, som er baseret på værdien af deres ejendom. Dette vil sige værdien af hele ejendommen inkl. den grund som boligen ligger på. For at kunne gøre dette laver den danske stat offentlige ejendomsvurderinger som disse skatter bliver baseret på. Det er derfor vigtigt at disse vurderinger er retvisende og ikke mindst forklarbare, således at en borger kan forstå hvilke parametre der ligger til grund for ejendomsvurderingen. Til dette project har jeg valgt at arbejde med anonymiseret data fra mit arbejde i udviklings- og forenklingsstyrrelsen, hvor jeg til dagligt arbejder med netop dette. Datasættet består af ejendomssalg fra en 6 årig periode. Ud over selve salgspriserne består data også af en lang række attributter som beskriver karakteristika ved selve boligen. Det kan f.eks. være tagmateriale, boligens opførelsesår, information om størrelsen af huset og grunden eller bbr koder som dækker over boligens anvendelse. Der ud over består data også af en lang række attributter som fortæller noget om hvor boligens beliggenhed. Det kan f.eks. være boligens koordinater, områdepriserne (baseret på de nærmeste nabosalg) eller information om afstanden til kyst og skov eller afstand til motorvej og jernbane. Data kommer fra en række forskellige registre og offentlige styrrelser som eks. BBR og Styrrelsen for Dataforsyning og Effektivisering.
Til dette projekt vil jeg overordnet set prøve at se hvor godt man kan forudsige ejendomsværdier ud fra salgspriserne fra en 6-årig periode.
Jeg vil med Principal Component Analysis få et overblik over de data der er til rådighed og få et visuelt overblik over attributterne. Herefter vil jeg med unsupervised learning forsøge at gruppere det data jeg har således at jeg ud fra det kan generere yderlige attributter som kan indgå i modellen. Jeg vil her specifikt prøve at se om det er muligt at gruppere salgene i forskellige boligtyper. Jeg vil i samme omgang også forsøge at frasorterer outliers i data med anomaly detection. Herefter vil jeg med regressions model forsøge at kaste lys

over projektets overordnede problem ved at forsøge at forudsige huspriserne ud fra salgspriser. I tilfælde af at modellen ikke ikke kan komme med en god prædiktion af en given ejendom vil det være muligt at denne ejendom bliver manuelt værdiansat af en sagsbehandler. Jeg vil derfor til slut med en Klassifikationsmodel forsøge at estimerer om en ejendom skal ud til manuel sagsbehandling baseret på dens estimerede ejendomsværdi.

1.2 Detaljeret beskrivelse af data

Det salgsdata som jeg har valgt at arbejde med dækker i udgangspunktet 305701 observationer med 122 attributter. Inden jeg går i gang med at kigge på data har jeg valgt at lave en oprydning i data. Mange af attributterne har ikke noget med selve ejendommen at gøre men er forretningsmæssige oplysninger som ikke er relevante for denne opgave. Desuden dækker observationerne mange forskellige typer af ejendomssalg. Det er en blanding af parcelhussalg, rækkehussalg, sommerhussalg, salg af ejerlejligheder mm. og ud fra et forretningsmæssigt perspektiv giver det ikke mening at træne en model på alle salg da ejendomstypen vil påvirke salgsprisen. F.eks. vil der på sommerhuse være restriktioner på hvor mange dage om året man må bo i sommerhuset og der kan være i sommerhusområder være andre regler for hvad man må bruge sin grund til end der er i et parcelhusområde. Jeg har derfor ligeledes valgt at reducerer antallet af observationer således at de kun dækker almindelige parcelhus. Dette er gjort ved kun at beholde alle de ejendomssalg, hvor ejendommen i BBR er registreret med enheds- og bygningsanvendelsen 120. Inden jeg i denne opgave anvender ejendomssalgene er deres salgspriser blevet fremskrevet til den sidste handelsdato. Det er de gjort med henblik på at neutralisere de prissvingninger som er i den 6 årige periode. Disse vil blive refereret til som de fremskrevne handelspriser.
Når alle disse grove datasorteringer er foretaget er der 241643 observationer tilbage og 106 attributter. Hertil kommer det at der er en del af attributterne som mangler værdier for en procentdel af det samlede antal observationer. Det er især i forhold til variable fra BBR, som beskriver forskellige karakteristika ved selve boligen. Her har jeg har valgt at fjerne alle de attributter som har mere end 95% manglende værdier. (se bilag)
Modellen der skal trænes skal som udgangsunkt kunne prædiktere værdien af et standard parcelhus. Data som modellen trænes på skal derfor også være salg af standard parcelhuse. Data er derfor blevet ensrettet på følgende måde:

- Antallet af værelser skal være større end 1 og mindre en 10.
- Boligarealet skal være større end 50 kvm og mindre end 500 kvm.
- Boligens alder skal være større end 0 men mindre end 100 år.
- Antallet af etager skal være større end 0 og mindre end 4.
- Antallet af badeværelser skal være større end 0 og mindre end 4.

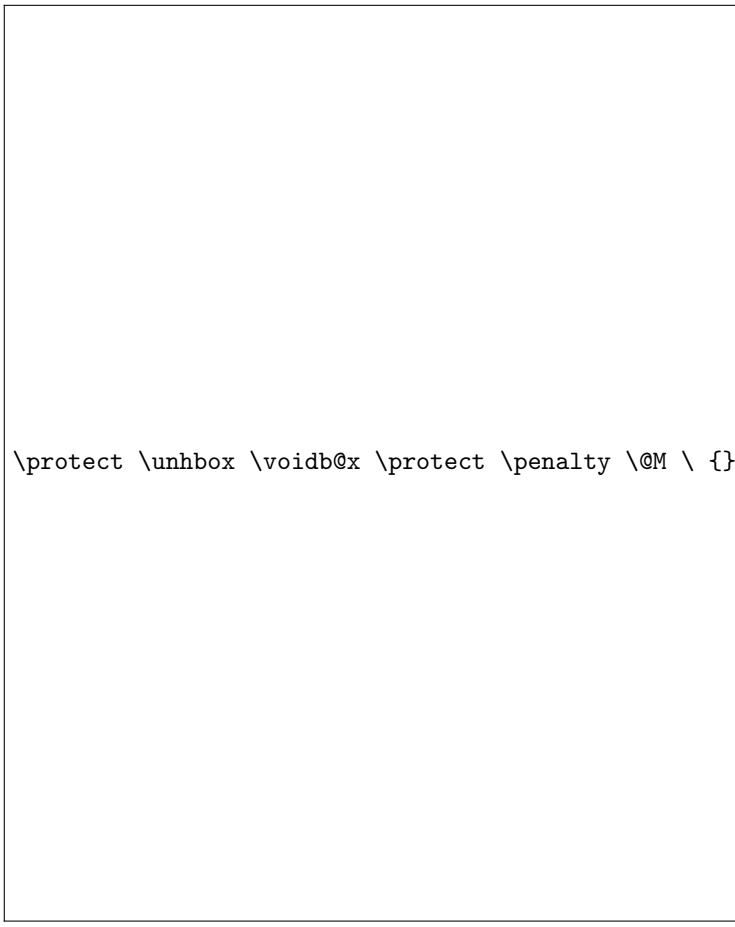
- Antallet af toiletter skal være større end 0 og mindre end 4.
- Den fremskrevne kvm-pris for salgene skal være større end 0 men mindre end 30.000 kr.

Slutteligt er der taget en forretningsmæssig beslutning om at udvælge de attributter som menes at have størst betydning i forhold til at forudsige værdien af et standard parcelhus. (se bilag).
De to attributter der dækker over tagtypematriale og ydervægsmateriale er diskrete variable som fordel kan normaliseres med en one-out-of-k transformering.
Afstand til kyst og afstand til motorvej er to variable som jeg har valgt at binariserer. Det er en beslutning som er blevet taget da data for disse to features forud for denne rapport er blevet imputeret. For afstand til kyst er afstanden op til 1500 meter målt. Alt herover er imputeret til 1501 meter. Ligeledes er gjort for afstand til motorvej. Ud fra et forretningsmæssig synspunkt har det afstand til kyst kun en påvirkning på ejendomsprisen hvis kysten ligger inden for omkring 300 meter. Ligeledes er det kun værdipåvirkende hvis en ejendom ligger inden for omkring 100 meter fra en motorvej. Valget med at binarisere disse variable gør at der bliver taget hånd om alle de imputerede værdier, men der bliver selvfølgelig samtidig tabt lidt information ved at gøre dette.

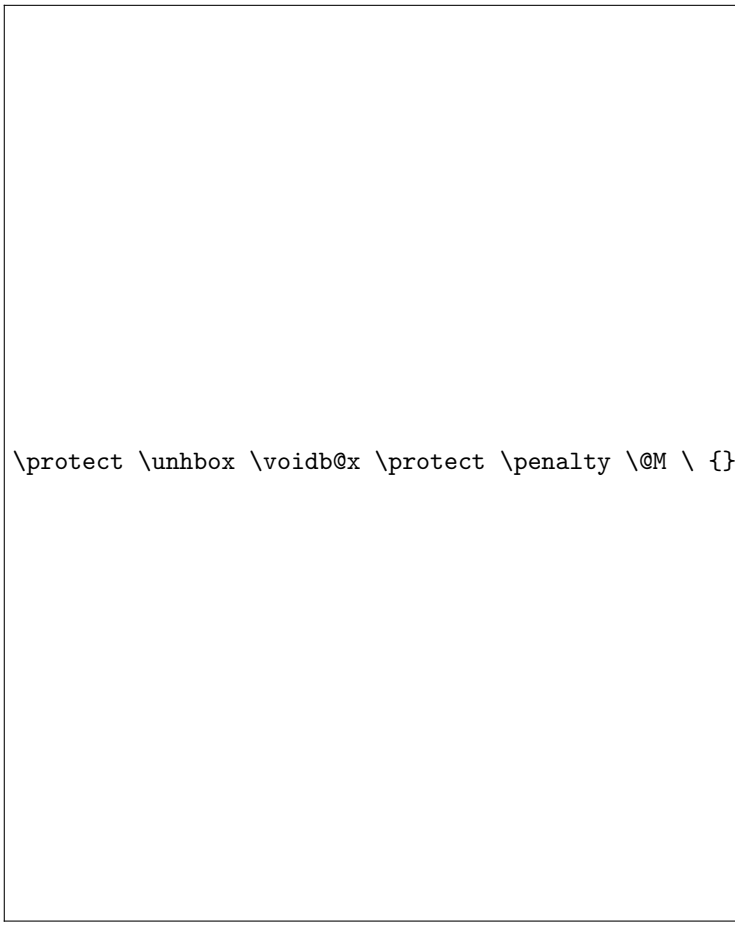
1.3 Data visualisering heriblandt Principal Component Analysis (PCA)

Principal Component Analysis (PCA) er en metode som kan bruges til at reducere dimensionerne data. Man kan have mange dimensioner data, men hvis de alle sammen er med til at forklare sammen tendens er det 'sande' antal af dimensioner lavere end antallet af attributter. Målet med at lave PCA er at reducere dimensionerne i data uden at reducere variationen, således at man ender op med data som med færre dimensioner, men uden at der tabes information. PCA fungerer kun ud fra antagelsen om at der er en linear forklaring i data med færre dimensioner. De bedste projektioner af data ned på et subspace er dem hvor observationer er spredt ud (høj varians), men samtidig hvor residualerne reduceres. Vektoren bliver valgt ud fra at den skal være en eigenvektor den datamatrice som har den højeste eigenværdi. Singular Value Decomposition (SVD) er en metode som for en hvilken som helst $N * M$ matrix udregner eigenvektoren med den højeste eigenværdi.
Ejendomsdata er blevet klargjort. Data er blevet tranformeret. Nogle variable er blevet tranformeret med one-out-of-K transformation, mens enkelte er blevet binariseret. Til PCA er det første trin at standardisere data, således at attributternes værdier er på samme skala. Selve standardiseringen består i at trække gennemsnittet fra hver attribut, hvorefter der også er blevet divideret med standardafvigelsen. For data betyder det at hver attribut reskaleres således

at de får et gennemsnit på 0 og en standardafvigelse på 1. Årsagen til at en reduktion af dimensionerne er ønskværdig er at det for nogle typer af algoritmer kan være med til at forøge deres nøjagtighed. Dette er eksempelvis tilfældet med xgboost algoritmen.
Efter alle datatransformationerne består data af 185018 observationer (N) med 23 features (M). Dette data skal senere danne grundlaget for regressionsanalysen, men inden da bliver der med en korrelationsanalyse og en PCA taget stilling til hvorvidt det er muligt at reducere demensionerne i data. Resultatet af korrelationsanalysen er vist i et korrelationsplot. Resultatet af Korrelationsanalysen viser at der er en stor positiv korrelation mellem den fremskrevne kvadratmeter pris og den vægtede gennemsnitspris for de nærmeste naboer. Der er desuden også an større positiv sammenhæng mellem antallet af værelser og boligarealet. Disse to positive sammenhænge giver logik rigtig god mening. Salgspriser er i høj grad styret af det område som ejendommen ligger i. Ligger ejendommen i et dyrt område, vil naboerne blive solgt til høje handelspriser og det samme vil højst sandsynligt også gælde for den specifikke ejendom. Samtidig vil der typisk også være flere væresler jo større boligareal en ejendommene har.



/Documents/projects/billeder/pca_plot1.pdf"



/Documents/projects/billeder/pca_plot2.pdf"

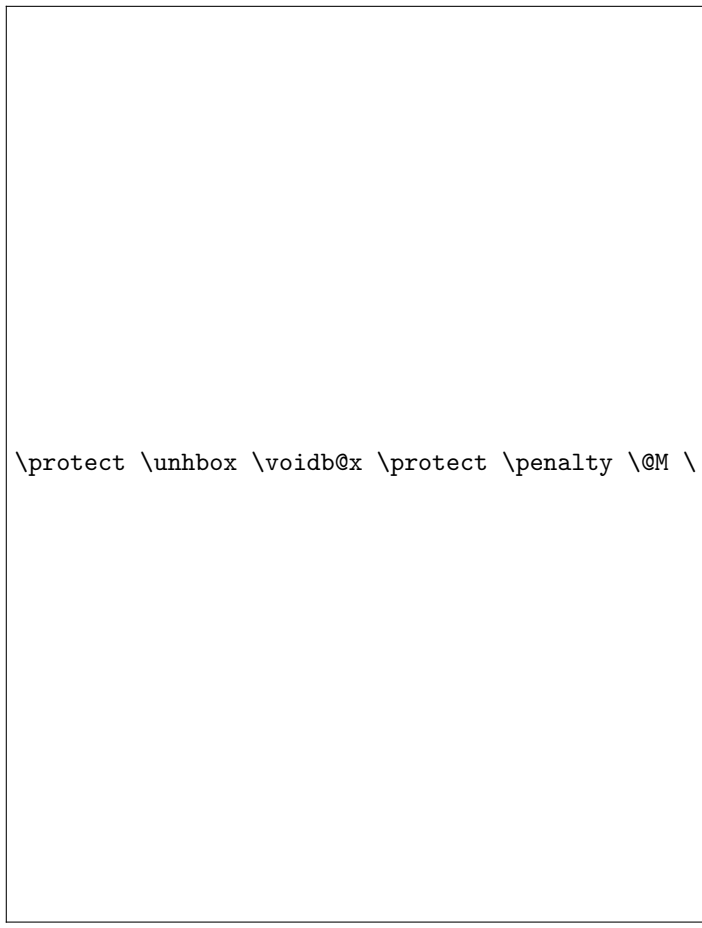
Som en del af PCA udregnes herefter Singular Value Decomposition (SVD). De første 16 principal components kan forklare 90% af variationen i data. For at kommme over 95% skal man have de 18 første komponenter. Ud af de i alt 23 mulige komponenter er det med dette data ikke muligt at reducerer mange komponenter væk uden også at miste variation i data.
Ved at have antallet af komponenter som er mindre end antallet af attributter i ens datasæt bliver information tabt, og hvorvidt man med fordel kan bruge PCA skal bestemmes ud fra den pågældende problemstilling. I den videre opgave har jeg valgt at gå videre med mit originale datasæt som det så ud før PCA.

Chapter 2

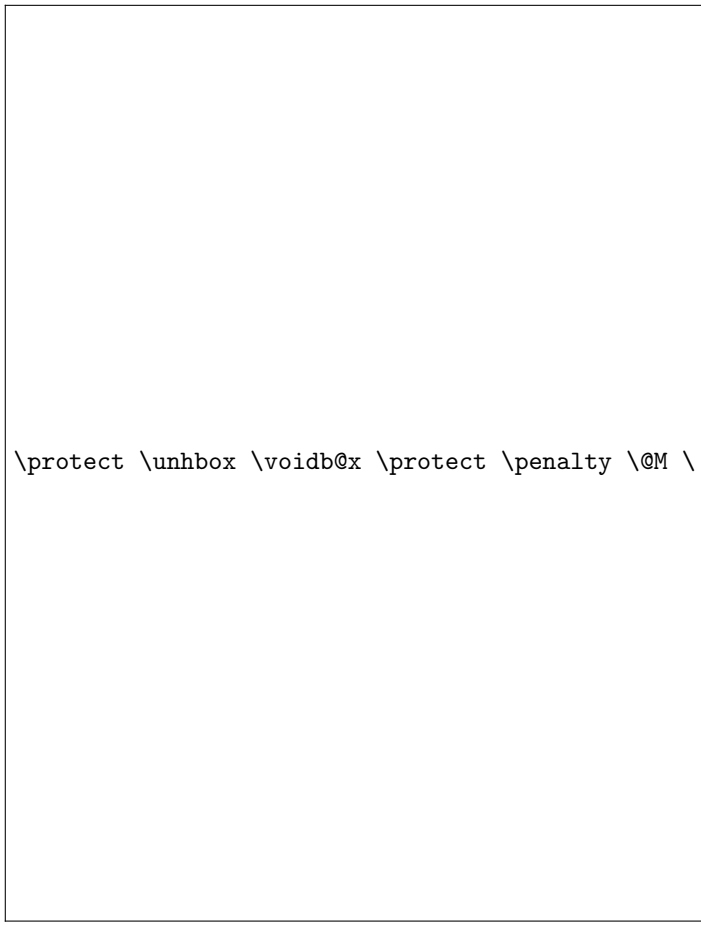
Part II- Supervised learning

2.1 Regression - part A

I 2 del er formålet at bruge det rensede data fra del 1 til at forudsige fremskrevne kvadratmeterpriser ud fra forskellige variable. Til dette formål vil jeg træne en lineær regressionsmodel, da den afhængige variabel "fremskreven kvadratmeterpris" er en numerisk variabel. I det tilfælde af at det havde været en diskret variabel ville en klassifikationsmodel være blevet brugt i stedet. Håbet med denne regressionsanalyse er at man ud fra relativt få variable og en relativt simpel model vil kunne forudsige ejendomspriserne. Forud for regressionsanalysen er data blevet tranformeret. For faktorvariablene tagtype og vægmateriale har jeg valgt at tranformere med en one-of-k transformering. Herefter er alle attributter blevet standardiseret, således at de har en gennemsnit på 0 og en standardafvigelse på 1.



/Documents/projects/billeder/linear_plot_1.pdf"



/Documents/projects/billeder/linear_plot_2.pdf"
I en multivariate lineær regressionsmodel kan man ikke på samme måde plotte den fittede model på det todimensionelle plot. Her kan man i stedet estimerer hvor godt modellen fitter til data ved at minimere summen af de kvadrerede afvigelser (RSS). Der findes flere forskellige typer af algoritmer hvis formål er at finde de parametre/vægte som laver det bedste fit til data ved at minimerer 'cost'.
Overfitting er når den trænede model er så god til at forudsige det data den er trænet på at den ikke kun fanger tendenserne i data men at den også fanger støj. For at undgå at den trænede model overfitter kan man bl.a. bruge metoden Krydsvalidering. Ved krydsvalidering opdeles data i to mindre datasæt. Et datasæt som der trænes på og et datasæt som modellen testes på. Selve opsplitningen i test og træingsdata kan gøres flere gange.
Når modellen er trænet kan man ud fra modellen koefficient estimater se hvor meget de enkelte parametre bidrager med hvis værdien for den pågældende parameter øges. I tilfælde af at de uafhængige variable som indgår i modellen er korrelerede kan estimaterne være svære at forklare, fordi noget af effekten indirekte ligger i den korrelerede parameter. Dette påvirker dog ikke

nødvendigvis selve prædiktionen.
I tilfælde at at ens variable ikke er er additive kan man med fordel kigge på interaktionerne mellem to variable. Dette er endnu en måde at få et bedre fit.