TTK4260 Innføring i Multivariat Datamodellering

Kristian Løvland

Innhold

1 Introduksjon

Dette dokumentet er et forsøk på å skaffe meg en strukturert oversikt over et middels strukturert fag. Ting er først og fremst hentet fra slides, men også noe fra lærebøker jeg kanskje legger til som kilder senere.

2 Grunnleggende statistikk

Til grunn for alle multivariate metoder som kommer senere, ligger den grunnleggende statistikken som forhåpentligvis er kjent fra før. Den oppsummeres her i korte trekk.

2.1 Kort om notasjon

Gjennom denne oppsummeringen brukes følgende notasjon

- Skalare variabler og funksjoner skrives med små bokstaver – $\alpha, x, g(\cdot)$
- Vektorvariabler og -funksjoner skrives med små, fete bokstaver – $\alpha, \mathbf{x}, \mathbf{g}(\cdot)$
- Matriser skrives med store bokstaver – A, X
- Transponert, invers, pseudoinvers – $\cdot^T, \cdot^{-1}, \cdot^\dagger$
- Definisjonsmengder skrives med kaligrafiske bokstaver, eller stor theta – $\mathcal{X}, \mathcal{Y}, \mathcal{D}$ eller Θ

2.2 Definisjoner

En funksjon $\phi(\boldsymbol{y}) : \mathcal{Y}^N \mapsto \mathbb{R}^M$ er en **observator** dersom den er målbar, og den er uavhengig av θ. En observator $\phi(\boldsymbol{y}) : \boldsymbol{y}^N \mapsto \Theta$ er en **estimator** dersom den er målbar og uavhengig av θ (dvs. den er en observator med verdimengde lik Θ.

2.3 Antagelser

Når vi snakker om regresjon, antar vi at dataen y_t genereres av funksjonen $y_t = f(u_t; \theta) + v_t$, og at dette resulterer i et datasett $\mathcal{D} = \{(u_t, y_t)\}_{t=1,\dots,N}$. Ved hjelp av parametre $\theta \in \Theta$, hypoteserommet vårt, vil vi finne en estimator $\widehat{\theta} \in \Theta$ som best mulig forklarer datasettet vårt.

Hva betyr det å "best mulig forklare $\mathcal{D}$? Det finnes det ulike tolkninger av.

2.4 Minste kvadrater

En naturlig tolkning av spørsmålet om å best mulig forklare datasettet vha θ, er å finne parameteren som ved bruk av vår antatte modell $f(u_t;\theta)$, gir den korteste avstanden fra predikert datasett til faktisk datasett. "Avstand" har her den vanlige, euklidiske tolkningen, slik at en minste kvadraters estimator av en parameter θ gitt et datasett $\mathcal{D}$ og en modell $f(u_t;\theta)$, er gitt av

$$\widehat{\theta}_{\mathrm{LS}} = \arg\min_{\theta\in\Theta} \left\| \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} - \begin{bmatrix} f\left(u_1;\theta\right) \\ \vdots \\ f\left(u_N;\theta\right) \end{bmatrix} \right\|^2 = \arg\min_{\theta\in\Theta} \sum_{t=1}^{N} \left(y_t - f\left(u_t;\theta\right)\right)^2 \tag{1}$$

En nyttig verdi som følger av dette estimatet er residualen $r_t(\theta) := y_t - f\left(u_t;\theta\right)$.

En nyttig klasse av problemer er de **separable** problemene. Disse er på formen

$$y_t = \sum_{j=1}^{n} \theta_j \phi_j\left(u_t\right) + e_t \tag{2}$$

Dvs. at parameterne som skal estimeres inngår lineært i modellen vår. Da kan $\phi(u_t)$ være så komplisert den vil, LS-problemet vil uansett reduseres til et lineært likningssett på formen

$$\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} \phi_1\left(u_1\right) & \cdots & \phi_n\left(u_1\right) \\ \vdots & & \vdots \\ \phi_1\left(u_N\right) & \cdots & \phi_n\left(u_N\right) \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_N \end{bmatrix} \tag{3}$$

Som mer kompakt kan skrives som

$$\boldsymbol{y} = \Phi(\boldsymbol{u})\boldsymbol{\theta} + \boldsymbol{e} \tag{4}$$

Målet vårt er å minimere $\mathbf{e}$. Dette kan gjøres vha. lineær programmering, men om vi ikke har begrensninger i θ, dvs. $\Theta = \mathbb{R}^n$ for en eller annen n, kan dette løses eksplisitt som

$$\widehat{\theta}_{\mathrm{LS}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^n} \|\boldsymbol{y} - \Phi(\boldsymbol{u})\boldsymbol{\theta}\|^2 \tag{5}$$

Ved å derivere dette og sette det lik null får vi **normallikningene**

$$\Phi(\boldsymbol{u})^T\Phi(\boldsymbol{u})\widehat{\theta}_{\mathrm{LS}} = \Phi(\boldsymbol{u})^T\boldsymbol{y} \tag{6}$$

Men hva om $\Phi(\boldsymbol{u})^T\Phi(\boldsymbol{u})$ ikke er invertibel? Da vil ikke normallikningene ha noen entydig løsning. Dette ordner vi ved å bruke **pseudoinversen**, som har noen kjekke egenskaper (nærmeste løsning hvis ingen løsning eksisterer, løsning med minst norm hvis mange løsninger eksisterer).

Noen ganger er imidlertid ikke alle tilstander like viktige. Dette løser vi ved å multiplisere hver feil med en vekt, slik at minimeringsproblemet (for ubegrensede problemer) blir

$$\widehat{\theta}_{\text{WLS}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^n}(\boldsymbol{y} - \Phi(\boldsymbol{u})\boldsymbol{\theta})^T W(\boldsymbol{y} - \Phi(\boldsymbol{u})\boldsymbol{\theta}) \tag{7}$$

Mer kompakt kan dette skrives som $= \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^n} \|\boldsymbol{y} - \Phi(\boldsymbol{u})\boldsymbol{\theta}\|^2_{W^{-1}}$. Et nytt sett med normalligninger faller ut av dette, nå blir

$$\Phi(\boldsymbol{u})^T W\Phi(\boldsymbol{u})\boldsymbol{\theta} = \Phi(\boldsymbol{u})^T W\boldsymbol{y} \tag{8}$$

som kan løses likt som tidligere, f.eks. med pseudoinvers.

Hva om problemet vårt er ulineært (ikke separabelt)? Optimeringsproblemet kan fortsatt være veldefinert, og da kan det løses numerisk. MATLAB gjør dette med `fmincon`, og de fleste programmeringsspråk med respekt for seg selv har rammeverk som gjør det samme.

2.5 Maksimal sannsynlighet

Forrige avsnitt ga en rent geometrisk tolkning av minste kvadrater. Ofte har min imidlertid kunnskap om de statistiske egenskapene til støyen og feilen som forsøpler dataen din, og denne informasjonen er gjerne nyttig å bruke. Utgangspunktet for dette er sannsynlighetsfordelingen, skrevet som $p(y;\theta)$. Ofte operer man med θ fiksert og y varierende. I vårt tilfelle er y gitt (den utgjør datasettet vårt $\mathcal{D}$, og vi er ute etter å finne en θ som best forklarer dette. Da kalles $p(y;\theta)$ for **sannsynlighet**.

Med dette definert er vi klare for å definere vår maksimale sannsynlighetsestimator

$$\widehat{\theta}_{\text{ML}} = \arg\max_{\theta\in\Theta} p(\mathcal{D};\theta) \tag{9}$$

Vi ser at denne ligner i formen på LS-estimatoren, men at funksjonen p gir oss mer valgfrihet, og muligheter til å inkludere kjent informasjon om modell og data.

Det er verdt å merke seg at et ML-estimat ikke nødvendigvis trenger å eksistere. Man kan komme opp med massevis av eksempler på at denne ikke

eksiterer, f.eks. kan man la Θ er en åpen mengde. Hvis p er kontinuerlig og Θ er kompakt tror jeg imidlertid vi kan føle oss ganske trygge, i hvert fall hvis man er av typen som stoler på det Weierstrass hadde å si oss.

Et viktig eksempel på en ML-estimator er når p er normalfordelt. Da vil sannsynlighetsfunksjonen til et datasett være gitt av

$$p\left(y_1, \ldots, y_N; m, \sigma^2\right) = \prod_{t=1}^{N} p\left(y_t; m, \sigma^2\right) = \prod_{t=1}^{N} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{\left(y_t - m\right)^2}{\sigma^2} \right) \right) \tag{10}$$

Dette grumsete uttrykket motiverer definisjonen av log-sannsynlighetsfunksjonen. Siden $\log(\cdot)$-funksjonen er monotont stigende i input-argumentet sitt, vil maksimerende input til funksjonen være lik maksimerende input til logaritmen av funksjonen. Vi definerer

$$\ell(\theta) := -\log p(\mathcal{D}; \theta) \tag{11}$$

I eksempelet med normalfordelt p vil vi nå kunne formulere

$$\widehat{\theta}_{\mathrm{ML}} = \arg\max_{\theta \in \Theta} p(\mathcal{D}; \theta) = \arg\min_{\theta \in \Theta} \ell(\theta) \tag{12}$$

Eksponential- og logaritmefunksjonen spiller hverandre gode, og ved litt regning kan man se at

$$\arg\min_{m \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+} \ell\left(m, \sigma^2\right) = \arg\min_{m \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+} N \log\left(\sigma^2\right) + \frac{\sum_{t=1}^{N} \left(y_t - m\right)^2}{\sigma^2} \tag{13}$$

Dennes gradient settes lik null, men her vil det inngå informasjon vi ikke har tilgang på. Vi ender opp med å benytte estimatene, og får

$$\bar{m} = \frac{1}{N} \sum_{t=1}^{N} y_t \tag{14}$$

$$\bar{\sigma}^2 = \frac{1}{N} \sum_{t=1}^{N} \left(y_t - \bar{m}\right)^2 \tag{15}$$

som ser fornuftig ut.

2.6 Maksimal a posteriori

Vi begynner med Bayes' lov

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{P(B|A)P(A)}{P(B)} \tag{16}$$

som kan bevises ved Venn-diagram eller lignende.

Ved å bytte ut A og B med θ og y, får vi en måte å oppdatere vår tro om en variabel sin fordeling på, basert på data. Vi er imidlertid avhengige av å ha en initiell formening om fordelingen til θ, dette kalles en **prior**. I praksis vil denne gjerne gjøre få antagelser, men utelukke fullstendig urealistiske muligheter (f.eks. utelukke negativ høyde på personer, om man vil estimere dette). Med en modell $P(y|\theta)$, en prior $P(\theta)$, og data som gir oss $P(y)$, får vi da

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)} \tag{17}$$

Basert på denne nye, evidensbaserte fordelingen, kan vi definere estimatoren

$$\widehat{\theta}_{\text{MAP}} := \arg\max_{\theta\in\Theta} P(\theta|y) = \arg\max_{\theta\in\Theta} \frac{P(y|\theta)P(\theta)}{P(y)} = \arg\max_{\theta\in\Theta} P(y|\theta)P(\theta) \tag{18}$$

som vil være moden til den posteriore fordeling. Merk at denne ikke trenger å være representativ for fordelingen, f.eks. hvis fordelingen har en smal, høy topp langt unna "tyngdepunktet".

2.7 Statistiske ytelsesindekser

Det nevnes tre bruksområder for indekser som måler statistisk ytelse

- Regresjonsproblemer
- Klassifiseringsproblemer
- Sammenligning av sannsynlighetsfordelinger

Vi går gjennom disse i den rekkefølgen.

2.7.1 Regresjonsproblemer

En naturlig indeks er **Mean Squared Error (MSE)**

$$\text{MSE} = \mathbb{E}\left[\|\theta - \widehat{\theta}\|^2\right] \tag{19}$$

Av denne følger **Root Mean Square Error**

$$\text{RMSE} = \sqrt{\mathbb{E}\left[\|\theta - \widehat{\theta}\|^2\right]} \tag{20}$$

Det er viktig å merke seg at siden MSE er en funksjon av θ, så kan den ikke regnes ut. Hvorfor bryr vi oss om den da? Tja, den kan i hvert fall inspirere lignende indekser. **Residual Sum of Squares (RSS)** baserer seg på residualene til estimatene

$$\text{RSS}(\widehat{\theta}) := \sum_i \left(y_i - \widehat{y}_i(\widehat{\theta})\right)^2 \tag{21}$$

Det finnes imidlertid problemer med alle disse. Først og fremst er det et problem at de er avhengig av mengden av og størrelsen på dataen man vurderer estimatene av. Man vektlegger å unngå avvik i estimatene fra store målinger mer enn små. En metode som fungerer noe bedre, uten å bruke normalisering, er å bruke 1-normen i stedet for kvadratet. Dette gjøres i **Mean Absolute Deviaton (MAD)**

$$\text{MAD} := \mathbb{E}[|y - \widehat{y}|] \tag{22}$$

En metode som bruker en form for normalisering er **Fraction of Variance Unexplained (FVU)**

$$\text{FVU}(\widehat{\theta}) := \frac{\text{RSS}(\widehat{\theta})}{\text{var}(y)} = \frac{\sum_i \left(y_i - \widehat{y}_i(\widehat{\theta})\right)^2}{\sum_i \left(y_i - \frac{1}{N}\sum_i y_i\right)^2} \tag{23}$$

Denne må tolkes med måte, siden hva som er en god forklaringgrad er veldig avhengig av hva slags felt man jobber i, og det konkrete bruksområdet. Dette er uansett en mye brukt indeks, men da i form av R^2 Denne tolkes som "andel av variansen i avhengig variable som er predikerbar fra de uavhengige variablene".

2.7.2 Klassifiseringsproblemer

Vi diskuterer her klassifisering i form av "ja/nei". Da kan man gjøre to typer feil: Falsk positiv (**Type 1**) og falsk negativ (**Type 2**). Figur **??** viser definisjonen på en del uttrykk som beskriver egenskapene til en klassifikator.

$$\text{accuracy} := \frac{\#\text{ of true positives} + \#\text{ of true negatives}}{\text{total }\#\text{ of instances}}$$

$$\text{precision} := \frac{\#\text{ of true positives}}{\#\text{ of true positives} + \#\text{ of false positives}} = \frac{\#\text{ of true positives}}{\text{total }\#\text{ of instances predicted as true}}$$

$$\text{sensitivity} := \frac{\#\text{ of true positives}}{\#\text{ of true positives} + \#\text{ of false negatives}} = \frac{\#\text{ of true positives}}{\text{total }\#\text{ of true instances}}$$

$$\text{specificity} := \frac{\#\text{ of true negatives}}{\#\text{ of true negatives} + \#\text{ of false positives}} = \frac{\#\text{ of true negatives}}{\text{total }\#\text{ of negative instances}}$$

Fig. 1: Egenskaper til klassifikator

Muntlig kan disse forklares som

- **Prevalens** – Hvor ofte opptrer ja-tilfellet i datasettet vårt?
- **Nøyaktighet** – Hvor ofte har klassifikatoren rett?
- **Feilklassifiseringsrate** – Hvor ofte tar klassifikatoren feil?
- **Presisjon** – Når klassifikatoren gjetter ja, hvor ofte er dette rett?
- **Falsk positiv-rate** – Når svaret er nei, hvor ofte gjetter klassifikatoren ja?
- **Sensitivitet** – Når svaret er ja, hvor ofte gjetter klassifikatoren ja?
- **Spesifitet** – Når svaret er nei, hvor ofter gjetter klassifikatoren nei?

Man kan også kombinere to av disse for å få **F1-score**

$$\text{F1-score} = 2\frac{\text{presisjon} \cdot \text{ sensitivitet}}{\text{presisjon} + \text{ sensitivitet}} \tag{24}$$

Om man vil finne ut hvor ulike to klassifikatorer er, kan man bruke **Kappa-koeffisient**. Dette forklares ikke mer, men det eksisterer.

2.7.3 Sammenligning av sannsynlighetsfordelinger

Her kan man bruke **Kullback-Leibler-divergens**.

2.8 Bias vs. varians

Dette er en avveining man ikke slipper unna når men bedriver estimering. La oss bruke MSE for å illustrere dette. La

$$\begin{aligned} \mathcal{V} &:= \widehat{\theta} - \mathbb{E}[\widehat{\theta}] \\ \mathcal{B} &:= \mathbb{E}[\widehat{\theta}] - \theta \end{aligned} \tag{25}$$

$$\begin{aligned} \mathbb{E}\left[\|\hat{\theta} - \theta\|^2\right] &= \mathbb{E}\left[\|\widehat{\theta} - \mathbb{E}[\widehat{\theta}] + \mathbb{E}[\widehat{\theta}] - \theta\|^2\right] \\ &= \mathbb{E}\left[\|\mathcal{V} + \mathcal{B}\|^2\right] \\ &= \mathbb{E}\left[(\mathcal{V} + \mathcal{B})^T(\mathcal{V} + \mathcal{B})\right] \\ &= \mathbb{E}\left[\|\mathcal{V}\|^2 + \|\mathcal{B}\|^2 + 2\mathcal{V}^T\mathcal{B}\right] \\ &= \mathbb{E}\left[\|\mathcal{V}\|^2\right] + \|\mathcal{B}\|^2 \end{aligned} \tag{26}$$

Denne avveiningen henger sammen med hvor komplisert man gjør forklaringsmodellen $f(u_t; \theta)$. Om man gjør den veldig komplisert vil man kunne følge dataen nøyaktig, men man vil være utsatt for at dette ikke lar seg generalisere til andre datasett **overfitting**. Dette svarer til lav bias, men stor varians. Om modellen er for enkel vil man få en enkel modell som generaliserer, men man vil også kunne unngå å beskrive viktig struktur i dataen. Dette er **underfitting**, og svarer til liten varians, men stor bias.

Det finnes flere metoder som forsøker å gjøre denne avveiningen. Noen av dem er

- Akaikes informasjonskriterium
- Det Bayesiske informasjonskriteriumet
- Minimum lengde-beskrivelse

3 Multivariat Dataanalyse

Nå begynner vi på ekte her.

3.1 Eksperimentdesign