Projekt – Modellering

Intro

Nedan följer några råd/rekommendationer hur man kan tänka kring datahantering och modellering när man arbetar med maskininlärning/Data minng.

Data

- Se till att förstå era variabler. Beskriv tydligt vad som är era förklarande variabler och vad som är er responsvariabel. Beskriv tydligt vad som är en observation i ert dataset.
- Vid behov: Om ni har kategoriska variabler som är obalanserade överväg att slå ihop kategorier (om det är rimligt), detta gäller både reponsvariabeln och förklarande variabler.

- Eventuellt: undersök univariata extremvärden och överväg om enstaka observationer bör uteslutas från vidare analys.
- Hantera saknade värden:
    - Oftast så är det enklast att utesluta de observationer som har saknade värden
    - Ibland är det rimligt bra att använda någon sorts imputering. Man bör bara försiktig då en dålig imputering kan påverka vidare modellering negativt.

- Dela in data i tränings-, validerings- och testdata.
- Vid klassificering: undersök att förhållandet mellan andelen i era klasser är ungefär samma i de olika dataseten som ni delat upp i. Det är ok att det är variationer, men man vill undvika stora avvikelser och att någon klass helt saknas i något av dessa dataset.
- Lägg testdata åt sidan till slutet när modelleringen är klar. Testdata används när vi valt en slutgiltig modell och vill skatta den modellens generaliserbarhet på ny liknande data
- Vid behov standardisera variabler som är kontinuerliga. Spara x_mean_train och x_sd_train för alla variabler som standardiseras.
    - Standardisera valideringsdata och testdata vid behov baserat på x_mean_train och x_sd_train för motsvarande variabler.
- Ibland kan det vara ok att först dela upp i tränings och testdata, och sen standardisera träningsdata. Efter det så delar man upp i träningsdata och valideringsdata. Detta kan ibland vara motiverat om man använder ett nästlat valideringsschema (se nedan under Valideringsdata).

- Gör visualiseringar och beskrivande statistik baserat på träningsdata.

- Om outliers hittas eller andra problematiska observationer hittas: om ni vill utesluta vissa observationer, så till att sätta upp tydliga regler för vad som är en outlier. Dessa regler kan senare användas på valideringsdata eller testdata vid behov. Det viktiga är dessa regler ska baseras på träningsdata och att man inte ändrar dessa senare.
- Man ska dock vara försiktig med att utesluta observationer från data. Ibland kan problem ovanliga/extrema observationer hanteras med kostnadsfunktioner som kan hantera sådana värden. Tex använda MAE istället för MSE kan vara ibland lämpligt om det finns extrema värden i ens kontinuerlig responsvariabel.

- Vid behov: Gör lämpliga transformationer av variablerna i träningsdata. Dessa bör motiveras utifrån kunskap från träningsdata. Sen gör ni motsvarande transformationer i valideringsdata och testdata innan ni ska använda dessa. Notera dock om att era valda transformationer använder information från variablerna själva så kan ni ta den informationen från motsvarande variabel i träningsdata.
    - Exempel 1: Variabel x_8 log-transfomeras i träningsdata x8_log = log(x_8). Gör på samma sätt i valideringsdata och träningsdata. (Inget speciellt i detta fall.)
    - Exempel 2: Variabel x_4 är kontinuerlig och ni har bestämt att den ska bli en kategorisk variabel med två klasser. Ni gör uppdelning i klasser baserat på om värdena är större än medianen av x_4 eller ej. Låt oss kalla det värdet för x4_median_train. När ni ska göra motsvarande transformation av x_4 i valideringsdata och testdata, använd då x4_median_train för att göra uppdelningen.

Modellering – Skatta modeller

Ni ska beskriva era "praktiska metod": hur ni applicerar era metoder/modeller på ert problem. Ni ska alltså har ett strukturerat arbetssätt och ni ska beskriva det övergripande här under Metoddelen i rapporten. Sen under Resultat visar ni resultatet av er applicerade metod.

- När man analyserar data är det naturligt att ha en initial fas där men explorativt utforskar data och olika metoder kopplat till problemet. Denna fas behöver man oftast inte skriva om under Metod. Notera: denna fas ska vara relativt kort i tid.
- Sen när ni fått ungefärlig bild av vad som ni behöver göras för att analysera er frågeställning så ska ni bestämma ett strukturerat upplägg på hur ni ska använda era metoder/modeller på ert problem.
- Hur ska data delas upp? (träning, validering, korsvalidering, test) Man bör ha tänkt på det innan, men ibland behöver man göra justeringar.
- **Vilka modeller/hyperparametrar/inställningar ska användas/göras? Detta är ofta en svår och viktig fråga.**

- Ofta behöver man fixera några inställningar/hyperparametrar (pga man behöver avgränsa arbetat och pga tidsbegränsningar)
- Andra inställningar/hyperparametrar vill man hitta så optimala värden för som möjligt (ofta med hjälp av korsvalidering/valideringsdata).
    - Kolla i litteraturen vilka hyperparameter som verkar vara mer viktiga för specifika modeller, dessa vill vi hitta bra värden på.
- Ange tydligt vilka hyperparametrar som ni fixerar och vilka som ni vill försöker optimera för olika metoder/modeller. De som ni optimerar: ange tydligt vilka olika hyperparametervärden som ni väljer mellan för varje specifik hyperparameter.
    - Exempel: "För KNN används korsvalidering med tio uppdelningar för att bestämma k. Följande följande värden på k testas: k = 5, 7, 9,… , 25. Det k som ger minst genomsnittligt MSE i korsvalideringen används."
- Om man ska göra en hyperparametersökning och behöver skatta många modeller, då är det ofta smidigt att koda det som en loop, där man loopar över olika hyperparametervärden. I loopen så sparar man lämpliga resultat för träning och validering i tex en vektor eller matris. Gör man detta så kan man skatta förhållandevis många modeller utan för mycket arbete. Sen är det enkelt att ta fram resultat för alla modeller man skattat och jämföra.

- Hur ska metoder/modellerna ska utvärderas och jämföras? Vilka utvärderingsmått? Plottar?
    - Klassificering: Här vill man kolla på övergripande mått och klassvisa mer detaljerade mått.
    - Regression: Här vill man kolla på övergripande mått och sen göra lämplig residualanalys
    - Olika mått/plottar för att avgöra effekten av olika förklarande variabler (då det är relevant)
    - Ofta är det bra med tydliga tabeller där man presenterar mått för olika modeller på träningsdata och valideringsdata.
- När ni utvärderar era modeller på valideringsdata så kan man använda plottar av valideringsdata vid behov. Om vi vill göra plottar på testdata så gör man det i samband med att testdata används på slutet.

- Notera: om ni ska göra korsvalidering så används träningsdata.
- Det är ok att revidera sitt praktiska upplägg vid behov efter att man har börjat med själva analysen (se till att uppdatera beskrivningen av upplägget)

Valideringsdata

Ofta är det bra att använda valideringsdata för att välja hyperparametrar. När ni har hittat den uppsättning av hyperparametrar som ger minst valideringsfel, då är det vanligt att skatta om modellen med träningsdata **och** valideringsdata med den bästa uppsättningen av hyperparametrar. Tanken är att om vi har valt bra hyperparametrar så har vi en rimlig nivå av regularisering, så då kan vi kan skatta modellen med mer data utan att riskera tydliga problem med överanpassning.

Ibland vill man skapa nya valideringsdata som används nästlat (alltså i flera nivåer) i modelleringen. Vi kan vilja skapa en ny valideringsmängd baserat på träningsdata som bara används av en specifik modellklass. Notera att såna mer avancerade valideringsscheman bara är aktuella om man har hyfsat mycket data från början.

Exempel:

Vi vill jämföra lasso regression med neutrala nätverk på ett dataset. Vi delar först upp data i tränings-, validerings- och testdata.

1. Lasso regression: Vi använder korsvalidering på träningsdata för att bestämma lambda-parametern. Vi erhåller den bästa lasso regressionsmodellen.
2. Neurala nätverk: Vi skapar en ny valideringsmängd baserat på träningsdata som vi använder för att specificera hyperparameter för nätverket. (I Keras kan detta göras med intern validering när vi skattar modellen.) Vi erhåller den bästa neurala nätverket.

Vi tar de bästa modellerna från steg 1 och 2 och undersöker vilken som presterar bäst på det ”vanliga” valideringsdata. Den modell som är bäst blir vår valda modell.

Vi tar den valda modellen och skattar kostnadsfunktionen på testdata för att undersöka modellens generaliserbarhet till ny liknande data.

Så här har vi två nivåer med valideringsdata. En nivå som bara användes inom modellklassen neurala nätverk och en nivå som användas generellt för att jämföra mellan modellklasser.