Föreläsning 2 - Variabelselektion och regularisering

Josef Wilzen

2020-08-17

Outline

Bias, varians, brus

$$y = f(x) + \varepsilon \qquad E[\varepsilon] = 0 \qquad V[\varepsilon] = \sigma^2$$

$$\hat{y} = \hat{f}(x_{test})$$

Förväntad test MSE:

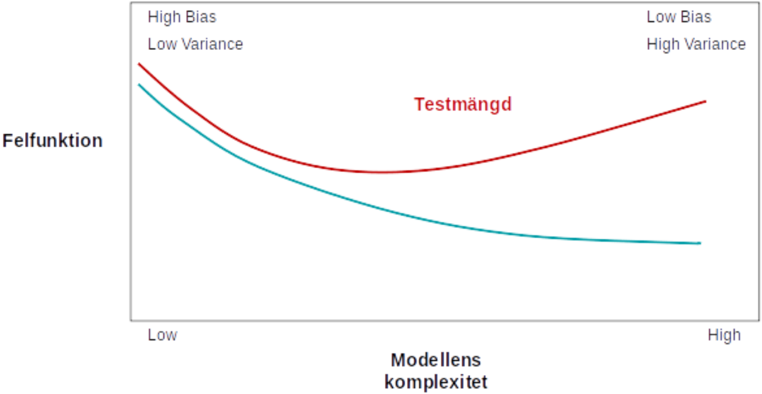
$$E\left[y_{test} - \hat{f}(x_{test})\right]^2 = V[\varepsilon] + V\left[\hat{f}(x_{test})\right] + Bias\left[\hat{f}(x_{test})\right]^2$$

- Brusvarians $V[\varepsilon]$: irreducibel brus
- Modellens varians $V\left[\hat{f}(x_{test})\right]$: Hur mycket kommer $\hat{f}$ att ändras när vi byter dataset
- Modellens skewhet $Bias\left[\hat{f}(x_{test})\right]$: Sytematisk skewhet eller modelleringsfel i modellen

Bias, varians, brus

Bias-variance-trade-off

- Vi fill ha
  - Lågt bias
  - Låg varians



Modellval

- Vi vill ha bra **generalisering** hos modellen
- Komplexa modeller överanpassar lätt
- "Komplexitet" betyder olika saker för olika modeller
  - Linjära modeller: fler variabler, interaktioner, transformationer av variabler
  - Neurala nätverk: bred och djup
  - Trädmodeller: djup

Regularisering

- Metoder för att undvika överanpassning
  - Hindra modellerna att bli för komplexa
  - Detta ger förhoppningsvis bättre generliseringsfel
  - **Mycket viktigt tema inom maskininlärning**
  - Betyder olika saker för olika metoder

Regression och klassificering

- Regression: y är kontinuerlig, bruset ε:
  - Antas ofta vara normalfördelat
  - Alt: t, gamma
- klassificering: y är kategorisk med 2 eller fler utfall
  - Binär: logistik/probit regression
  - Fler klasser: multinomial logistik/probit regression
  - Fler metoder senare i kursen

Förväxlingsmatris

		Predikterad klass	
		Class = 1	Class = 0
Sann klass	Class = 1	f_{11}	f_{10}
	Class = 0	f_{01}	f_{00}

- Precision:

$$P = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

- Felkvot (error rate):

$$E = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

Specificitet och sensitivitet

		Predikterad klass	
		Class = 1	Class = 0
Sann klass	Class = 1	f_{11}	f_{10}
	Class = 0	f_{01}	f_{00}

- Sensitivitet:

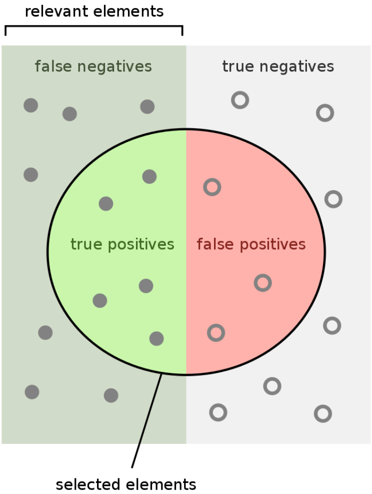
$$= \frac{f_{11}}{f_{11} + f_{01}}$$

- Specificitet:

$$= \frac{f_{00}}{f_{10} + f_{00}}$$

- Dessa mått är klasspecifika. Dessa formler betecknar klass 1.

Specificitet och sensitivitet



- By FeanDoe. Modified version from Walber's Precision and Recall https://commons.wikimedia.org/wiki/File:Precisionrecall.svg

Generaliserade linjära modeller

Generaliserade linjära modeller (GLM):

- $y_1, y_2, \ldots y_n$: oberoende från sannolikhetsfördelning från exponentialfamiljen
- Linjär prediktor: $X\beta$
- Länkfunktion som kopplar den linjära prediktorn till medelvärdet μ

$$g(\mu) = X\beta$$

Generaliserade linjära modeller

- Generaliserar linjär regression till andra reponsvariabler
  - Andra likelihoodfunktioner
- Reponsvariabler
  - Kontinuerlig: normal
  - Binär: Logistisk regression
  - Nomiell: Multinomiell logistisk regession
  - Frekvensdata: Poission regression

Linjär regression

- Likelihood: Normal
- Länkfunktion: identitetsfunktionen
- Skattas med genom att minimera:

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2$$

Logistisk regression

- Anta binär y. $p = P(y = 1)$
- Odds: $0 \leq \frac{p}{1-p} < \infty$, Log odds = logit: $-\infty < log\left(\frac{p}{1-p}\right) < \infty$
- Logistisk regression antar att log oddset för $P(y = 1)$ beror linjärt på de förklarande variablerna

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p \Leftrightarrow$$

$$p = \frac{1}{1 + exp(-(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p))}$$

$$= \frac{exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}{1 + exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}$$

Logistisk regression

- Logistiska funktionen:

$$p = \frac{1}{1 + exp(-(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p))}$$

- Prediktion: Om $P(y = 1|X_{test}, \beta) > 0.5 \rightarrow$ klass 1
- Skattas med maximum likelihood (ML), hitta β som maximerar:

$$l(\beta) = \prod_{i:y_i=1} P(y_i = 1|X_i, \beta) \prod_{i':y_{i'}=0} (1 - P(y_{i'} = 1|X_{i'}, \beta))$$

Multinomiell regression

- Anta att y kan anta K olika värden
- Multinomiell logistisk regression använder länkfunktionen:

$$P\left(y_i = k\right) = \frac{exp\left(\beta_{0,k} + \beta_k^T x\right)}{\sum_{l=1}^{K} exp\left(\beta_{0,l} + \beta_l^T x\right)}$$

- Kallas för softmaxfunktionen
- Notera att vi har: $\beta_{0,l} + \beta_l^T x$ för varje klass l. β är en matris med storlek $p \times K$
- Skattas med ML.

Modelval för linjär regression

Utgå från: y kontiuerlig med normal likelihood
Vi har ett antal förklarande variabler $X = (x_1, \ldots x_p)$. Vill vill hitta den delmängd som ger minst generaliseringsfel på testdata. Två alternativ:

- Välj ut en delmängd av variablerna och skatta med OLS
  - Best subset, Forward selection, Backward selection
- Behåll alla variablerna med sätt begränsningar på variablernas parameterrum (support) $\rightarrow$ det ger mindre flexibilitet $\rightarrow$ minskar risken för överanpassning
  - Ridge, lasso

Best subset

Notera det finns 2^p modeller att undersöka! Ex: $2^{20} = 1\,048\,576$

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Från "An Introduction to Statistical Learning with Applications in R" av Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

Forward selection

Algorithm 6.2 *Forward stepwise selection*

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Från "An Introduction to Statistical Learning with Applications in R" av Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

Backward selection

Algorithm 6.3 *Backward stepwise selection*

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Från "An Introduction to Statistical Learning with Applications in R" av Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

Utväderingsmått

- Indirekt skatta testfelet
  - Utgår från träningsmängden
  - Försöker minska den bias som uppstår när vi bara använder oss av träningsmängden och inte "all data"
- Direkt skatta testfelet
  - valideringsdata
  - korsvalidering

Indirekt skatta testfelet

$$C_p = \frac{1}{n}\left(RSS + 2d\hat{\sigma}^2\right)$$

$\hat{\sigma}^2$: skattas ofta från den fulla modellen
Straffar med: $2d\hat{\sigma}^2$
Litet $C_p \rightarrow$ litet testfel

$$\textit{adjusted}\, R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)}$$

Stort *adjusted* $R^2 \rightarrow$ litet testfel

Indirekt skatta testfelet

AIC,BIC och HQIC: based on ML skattning av modeller. Låt $log\left(\hat{L}\right)$ vara värdet på log-likelihoodfunktionen för optimla parametervärden.

$$AIC = 2k - 2log\left(\hat{L}\right)$$

$$BIC = k \cdot log\left(n\right) - 2log\left(\hat{L}\right)$$

$$HQIC = k \cdot log\left(log\left(n\right)\right) - 2log\left(\hat{L}\right)$$

Låga värden $\rightarrow$ litet testfel

Indirekt skatta testfelet

För linjär regression

$$AIC = \frac{1}{n\hat{\sigma}^2}\left(RSS + 2d\hat{\sigma}^2\right)$$

$$BIC = \frac{1}{n\hat{\sigma}^2}\left(RSS + log\left(n\right)d\hat{\sigma}^2\right)$$

$$HQIC = \frac{1}{n\hat{\sigma}^2}\left(RSS + log\left(log\left(n\right)\right)d\hat{\sigma}^2\right)$$

Linjär regression: $C_p \propto AIC$
Hur hög kostnad för fler parametrar?

- $n = 8$
$$2 < log\left(n\right)$$
- $n = 2000$
$$2 < log\left(log\left(n\right)\right) < log\left(n\right)$$
- BIC och HQIC väljer mindre antal parameterar generellt. För stora datamängder så tenderar AIC att välja för stora modeller.

Shrinkage

Shrinkage (krympning)

- Hindra parameterarna från att bli “för stora”
  - Sätta begränsningar på parameterrummet
  - Ändra deras support (värdemängd)

Vanligaste metoderna:

- Ridge: l^2-norm (Euclidean norm)
- LASSO: l^1-norm
- Standardisera era förklarande variabler först!

Ridge regession

- Vanlig regression minimerar konstandsfunktionen

$$f(\beta) = RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2$$

- Ridge= OLS + l^2-norm på β:

$$\arg\min_{\beta} f(\beta) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \qquad \lambda \geq 0$$

- λ är en **hyperparameter**.
- $\lambda \sum_{j=1}^{p} \beta_j^2$ = shrinkage penalty. Absolut stora värden på β "kostar mer". Hur mycket mer beror på λ (tänk prissättningen)
- Vad händer när $\lambda \to 0$? eller $\lambda \to \infty$?
- Notera att β_0 inte påverkas av konstandsfunktionen

Ridge regession

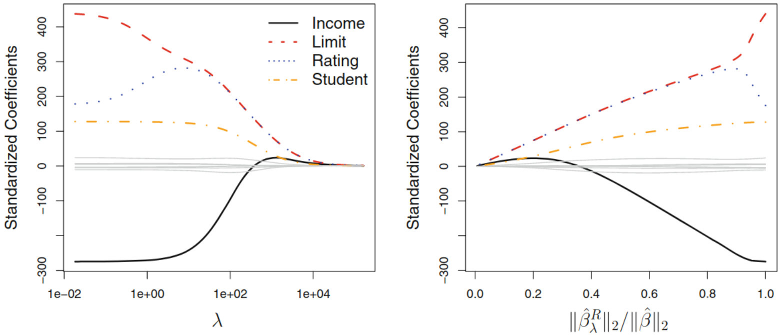


FIGURE 6.4. *The standardized ridge regression coefficients are displayed for the* Credit *data set, as a function of λ and $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$.*

Från "An Introduction to Statistical Learning with Applications in R" av Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

Ridge regession

- $\lambda \to 0$, $\hat{\beta}_{ridge} \to \hat{\beta}_{OLS}$
- $\lambda \to \infty$, $\hat{\beta}_{ridge} \to 0$
- $\sum_{j=1}^{p} \beta_j^2 = ||\beta||_2^2$: kvadrerad l^2-norm
- $\sum_{j=1}^{p} \beta_j^2$: definerar en hypersfär i det p-dimensionella rummet.
  - $p = 2$: cirkel, $p = 3$: sfär
- Krymper: "alla lika mycket"
  - två korrelerade variabler: tar med båda, men krymper effektstorleken

Lasso regression

- Vanlig regression minimerar konstandsfunktionen

$$f(\beta) = RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2$$

- lasso= OLS + l^1-norm på β:

$$\arg\min_{\beta} f(\beta) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \qquad \lambda \geq 0$$

- λ är en **hyperparameter**.
- $\lambda \sum_{j=1}^{p} \beta_j^2$ = shrinkage penalty. Absolut stora värden på β "kostar mer".
- lasso = least absolute shrinkage and selection operator

Lasso regression

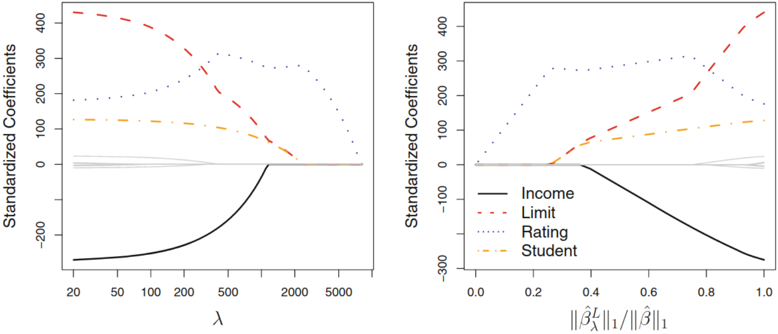


FIGURE 6.6. *The standardized lasso coefficients on the* Credit *data set are shown as a function of* λ *and* $\|\hat{\beta}^L_\lambda\|_1/\|\hat{\beta}\|_1$.

Från "An Introduction to Statistical Learning with Applications in R" av Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

Lasso regression

- $\lambda \to 0$, $\hat{\beta}_{lasso} \to \hat{\beta}_{OLS}$
- $\lambda \to \infty$, $\hat{\beta}_{lasso} \to 0$
- $\sum_{j=1}^{p} |\beta_j| = ||\beta||_1$: l^1-norm
- $\sum_{j=1}^{p} |\beta_j|$: definerar en polytop i det p-dimensionella rummet.
  - $p = 2$: diamant, $p = 3$: polyeder
- Lasso: kan tvinga vissa β av bli 0 $\to$ variabelselektion, glesa (sparse) lösningar
  - ex: Låt $p = 100$, Då kan lasso göra att endast 20 β-koefficienter är $\neq 0$
- Två korrelerade variabler:
  - "tar en och kastar bort den andra" $\to$ kan ge hög varians för testdata

Ridge och lasso

Kan formuleras om:

- Minimmera RSS med vilkoret
  - Ridge: $\sum_{j=1}^{p} \beta_j^2 \leq s$
  - Lasso: $\sum_{j=1}^{p} |\beta_j| \leq s$
  - Varje λ motsvarar ett s

Ridge och lasso

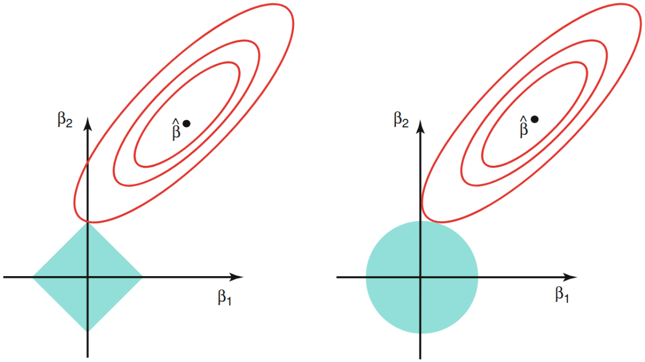


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*

Från "An Introduction to Statistical Learning with Applications in R" av Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

Rigde vs lasso

- Båda skattas enkelt med `glmnet()` eller `cv.glmnet()`.
- Vilken som är "bäst" beror på kontexten
- Ridge passar för när y beror på de flesta förklarande variablerna
  - Många variabler och ungefär samma effektstorlek
- Lasso passar för när y beror på de ett fåtal förklarande variabler
  - Fåtal variabler med hög effektstorlek, resten nära 0
- Lasso kan ofta vara lättre att tolka
- Lättre att göra inferens för parametrarna för ridge
- Ofta får frågan avgöras empiriskt för specifika dataset

Elasticnet regression

- Elasticnet kombinerar ridge och lasso
- Två hyperparametrar: λ och α

$$f(\beta) = RSS + \lambda \left[(1-\alpha) \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j| \right] \qquad \lambda \geq 0, 0 \leq \alpha \leq 1$$

- λ: generella styrkan på kostnaden
- α: relativa vikten på lasso och ridge, $\alpha = 1$ ger lasso.

Elasticnet regression

- Kan tvinga vissa β av bli 0 $\rightarrow$ variabelselektion, dock inte lika mycket som lasso
- Klarar av korrelerade/grupper av variabler bättre än lasso
- Nackdel: två hyperparameterar att specificera

Regularisering

- Överanpassning är stort problem inom maskininlärning
- Regularisering:
  - Förekommer mycket ofta
  - Kan se olika ut beroende på modellklassen
- Ridge och Lasso
  - $\sum_{j=1}^{p} \beta_j^2 = ||\beta||_2^2$: kvadrerad l^2-norm
  - $\sum_{j=1}^{p} \left|\beta_j\right| = ||\beta||_1$: l^1-norm
  - Förekommer i många olika varianter

Avslut

- Frågor? Kommentarer?
- Kurshemsidan
- Labben