MA8701 Advanced methods in statistical inference and learning

~~L3-4~~: Model selection and assessment

L3

Mette Langaas

~~1/15/23~~ 16.01.2023

Remember:

- 3 members to reference group! Philip, Didrik & Jacob.
- Office hours Monday and Friday at 9-10, 1236. → ok! Welcome!

Model assessment and selection

(ESL Ch 7.1-7.6,7.10-7.12)

The generalization performance of $\hat{f}$ can be evaluated from the EPE (expected prediction error) on an independent test set.

We use this for

- Model assessment: evaluate the performance of a selected model
- Model selection: select the best model for a specific task - among a set of models

Plan

1) Look at EPE(x_0) (now called Err(x_0)) and how model complexity can be broken down into irreducible error, squared bias and variance (should be known from before)
2) Study EPE (Err) unconditional and conditional on the training set

TODAY L3

3) Study optimism of the training error rate, and how in-sample error may shed light

L4

4) Cross-validation and .632 bootstrap estimates of EPE
5) How will we build on this in the rest of the course?

We finished 1) and 2) in L2, now we continue!

OPTIMISM OF THE TRAINING ERROR RATE

$\hat{f}(X)$ predictor for Y , (X,Y) random variables from $p(x,y)$

$\hat{G}$ $\qquad\qquad\qquad\qquad G$

training set $\mathcal{T} = \{(x_1,y_1),\dots,(x_N,y_N)\}$

(X,Y) $\qquad$ (X^0,Y^0) a new test point drawn from (X,Y)

Expected prediction error EPE (Err)

$$Err = \underset{X,Y}{E}\left(L(Y,\hat{f}(X))\right) = \underset{\mathcal{T}}{E}\Big[\underbrace{\underset{X^0Y^0}{E}\left(L(Y,\hat{f}(X))\,|\,\mathcal{T}\right)}_{Err_{\mathcal{T}}}\Big]$$

generalization error when $\mathcal{T}$ is kept fixed

We saw last time that $\overline{err} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$ is typically less than Err_{τ}. [Exercise 2.9]

Still hard to work with Err_{τ}, but it turns out to be easier if we fix the new observations (x^0) to be at the training set x_i's.

$$Err_{in} = \frac{1}{N}\sum_{i=1}^{N} E_{Y^0}\left[L(Y_i^0, \hat{f}(x_i)) \mid \tau)\right]$$

in sample error

new obs Y_i^0 at x_i

How does this compare to $\overline{err}$? Optimism op

$$op \equiv Err_{in} - \overline{err} = \frac{1}{N}\sum_{i=1}^{N} E_{Y^0}(L(Y_i^0, \hat{f}(x_i))T) - \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$

Is op positive or negative? Positive

Again hard to estimate – but possible to estimate ω

$\omega \equiv E_y(op)$

for squared loss: $\omega = \frac{2}{N} \sum_{i=1}^{N} \mathrm{cov}(\hat{y}_i, y_i)$

0-1 loss $\hat{y}_i = \{0,1\}$

entropy loss $\hat{y}_i \in [0,1]$

$f(x_i)$

$$E_y(Err_{in}) = E_y(\overline{err}) + \frac{2}{N} \sum_{i=1}^{N} \mathrm{cov}(\hat{y}_i, y_i)$$

Covariance result

For squared error, 0-1 loss, and "other loss functions" it can be shown

$$\omega = \frac{2}{N} \sum_{i=1}^{N} \mathrm{Cov}(\hat{y}_i, y_i)$$

Group discussion

1) Give an interpretation of the result.
2) How do you think this result can be used?
3) Study the derivation of the covariance formula for squared loss. This is Exercise 7.4 and solutions are available here and in the ESL solutions to exercises.

Expected in-sample prediction error

$$\mathsf{E}_{\mathbf{y}}(\mathsf{Err}_{\mathsf{in}}) = \mathsf{E}_{\mathbf{y}}(\overline{\mathsf{err}}) + \frac{2}{N}\sum_{i=1}^{N} \mathsf{Cov}(\hat{y}_i, y_i)$$

This is the starting point for several methods to “penalize” fitting complex models!

Look at ω for MLR (multiple linear regr.) with minimizing squared error

$$Y = f(X) + \varepsilon \qquad E(\varepsilon),\ Var(\varepsilon) = \sigma^2_\varepsilon$$

and in addition a linear fit in d inputs (ESL Ex 7.1)

We know that a linear fit has $Y = \mathbf{X}\beta + \varepsilon$, $\quad \hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y$

Full vector of predictions $\hat{Y} = X\hat{\beta} = \underbrace{X(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T}_{H} Y = \underset{N\times N}{H}\,\underset{N\times 1}{Y}$

our hat matrix — this is a so called linear smoother

If we find $\underset{N\times N}{Cov(\underset{N\times 1}{\hat{Y}}, \underset{N\times 1}{Y})}$ this is an $N\times N$ matrix and

the trace of this matrix will give us $\sum_{i=1}^{N} Cov(\hat{y}_i, y_i)$

sum of the diagonal elements

$$\mathrm{Cov}(\hat{Y}, Y) = E\left((\hat{Y}-E(\hat{Y}))(Y-E(Y))^T\right)$$

$$\mathrm{Cov}(HY, Y) = H\,\underbrace{\mathrm{Cov}(Y,Y)}_{\sigma_\varepsilon^2 I}$$

tr = trace

$$\text{Thus } \sum_{i=1}^{N} \mathrm{Cov}(\hat{y}_i, y_i) = \mathrm{trace}(\mathrm{Cov}(\hat{Y},Y)) = \mathrm{tr}(H\cdot \sigma_\varepsilon^2 I)$$

$\mathrm{trace}(kA) = k\,\mathrm{trace}(A)$

$$= \sigma_\varepsilon^2\,\mathrm{tr}(H) = \sigma_\varepsilon^2\,\mathrm{tr}(\underbrace{X}_{n\times d}(X^TX)^{-1}\underbrace{X^T}_{d\times n})$$

$$= \sigma_\varepsilon^2\,\mathrm{tr}(\underbrace{X^TX(X^TX)^{-1}}_{\substack{I \\ d\times d}}) = \underline{\underline{d\,\sigma_\varepsilon^2}}$$

$\mathrm{trace}(ABC) = \mathrm{trace}(CAB)$
invariant under
cyclic permutation

Result for ω

Additive error model and squared loss: $Y = f(X) + \varepsilon$, with $\hat{y}_i$ obtained by a linear fit with d inputs (or basis functions)

$$\omega = 2\frac{d}{N}\sigma_\varepsilon^2$$

Proof: ESL 7.1

Group discussion

- Comment on the derivation of ω - anything unclear?
- How does d and N and σ_ε^2 influence the average optimism?

$$\omega = \frac{2d}{N} \sigma_\varepsilon^2$$

- increase with d and σ_ε^2
- decrease with N

Three ways to perform model selection

- Estimate of expected in-sample prediction error (ESL Ch 7.5-7.6): We may develop the average optimism for a class of models that are linear in the parameters (Mallows Cp, AIC, BIC, ...) - and compare models of different complexity using $\mathrm{E}_{\mathbf{y}}(\mathrm{Err}_{\mathrm{in}})$. Remark: in-sample error is not of interest, but used to choose between models effectively.
- Estimate Err (ESL Ch 7.10-7.11): We may instead use resampling methods (cross-validation and bootstrapping) to estimate Err directly (and use that for model selection and assessment).
- In the data rich approach: we have so much data that we use a separate validation set for model selection (and a separate test set for model assessment). That is not the focus of ESL Ch 7.

Estimates of (expected) in-sample prediction error

We have the following result:

$$\mathsf{E}_{\mathbf{y}}(\mathsf{Err}_{\mathsf{in}}) = \mathsf{E}_{\mathbf{y}}(\overline{\mathsf{err}}) + \frac{2}{N}\sum_{i=1}^{N} \mathsf{Cov}(\hat{y}_i, y_i)$$

where now

$$\omega = \frac{2}{N}\sum_{i=1}^{N} \mathsf{Cov}(\hat{y}_i, y_i)$$

We now want to get an estimate of the average optimism, to get an estimate of the in-sample prediction error:

$$\widehat{\mathsf{Err}_{\mathsf{in}}} = \overline{\mathsf{err}} + \hat{\omega}$$

Comment: observe that $\overline{\mathsf{err}}$ is now an estimate of $\mathsf{E}_{\mathbf{y}}(\overline{\mathsf{err}})$ and even though we write $\widehat{\mathsf{Err}_{\mathsf{in}}}$ we are aiming to estimate $\mathsf{E}_{\mathbf{y}}(\mathsf{Err}_{\mathsf{in}})$. Focus now is on $\hat{\omega}$!

C_p statistics

for squared error loss (follows directly from the ω-result for additive error model)

$$C_p = \overline{\text{err}} + 2\frac{d}{N}\hat{\sigma}_\varepsilon^2$$

where $\hat{\sigma}_\varepsilon^2$ is estimated from a "low-bias model" (in MLR we use a "full model").
(This method is presented both in TMA4267 and TMA4268, see also exam question Problem 3 in TMA4267 in 2015 and solutions.)

Akaike information criterion (AIC)

Based on different asymptotic ($N \to \infty$) relationship for log-likelihood loss functions

$$-2\mathsf{E}[\log P_{\hat{\theta}}(Y)] \approx -\frac{2}{N}\mathsf{E}[\text{loglik}] + 2\frac{d}{N}$$

- $P_{\hat{\theta}}(Y)$: family of density for Y where the true density is included
- $\hat{\theta}$: MLE of θ
- loglik: maximized log-likelihood $\sum_{i=1}^{N} \log P_{\hat{\theta}}(y_i)$

Logistic regression with binomial loglikelihood

$$\text{AIC} = -\frac{2}{N}\text{loglik} + 2\frac{d}{N}$$

Multiple linear regression if variance $\sigma_\varepsilon^2 = \hat{\sigma}_\varepsilon^2$ assumed known then AIC is equivalent to C_p.

For nonlinear or similar models then d is replaced by some measure of model complexity.

AIC as function of tuning parameter (back to squared error loss)

We have a set of models $f_\alpha(x)$ indexed by some tuning parameter α.

$$\mathrm{AIC}(\alpha) = \overline{\mathrm{err}}(\alpha) + 2\frac{d(\alpha)}{N}\hat{\sigma}_\varepsilon^2$$

- $\overline{\mathrm{err}}(\alpha)$: training error
- $d(\alpha)$ number of parameters
- $\hat{\sigma}_\varepsilon^2$ estimated variance of large model

The model complexity α is chosen to minimize $\mathrm{AIC}(\alpha)$.

This is not true if the models are chosen adaptively (for example basis functions) this formula underestimates the optimism - and we may regard this as the *effective number of parameters* is larger than d.

Expected in-sample prediction error for binary classification

(Efron and Hastie (2016) page 225)
Misclassification loss function: $L(\hat{G}(X), G) = 1$ for incorrect classification and 0 for correct.
The training error is then $\overline{\text{err}} = (\#(\hat{G}_i \neq G_i))/N$.
The insample error is then $\frac{1}{N}\sum_{i=1}^{N} P(G_{0i}(X_i) \neq \hat{G}(X_i))$.
The estimate of (expected) in-sample prediction error is then

$$\widehat{\text{Err}_{in}} = \frac{\#(\hat{G}_i \neq G_i)}{N} + \frac{2}{N}\sum_{i=1}^{N} \text{Cov}(\hat{G}(X_i), G(X_i))$$

Here it is not "d"

where

$$\text{Cov}(\hat{G}(X_i), G(X_i)) = \text{E}(\hat{G}(X_i) \cdot G(X_i)) - \text{E}(\hat{G}(X_i)) \cdot \text{E}(G(X_i))$$

$$= \mu_i(1-\mu_i)[P(\hat{G}(X_i) = 1 \mid G(X_i) = 1) - P(\hat{G}(X_i) = 1 \mid G(X_i) = 0)]$$

where $\mu_i = P(G(X_i) = 1)$.

We can do model selection with only the training data! $E_y(Err_{in})$ || est.

Group discussion

What is the take home message from this part on "Estimates of (expected) in-sample prediction error"?

When is the Cov-result valid? squared loss, 0-1-loss and "some other loss functions"

THE EFFECTIVE NUMBER OF PARAMETERS (ESL 7.6)

MLR by quadratic loss gives $\hat{Y} = \underbrace{X(X^TX)^{-1}X^T}_{H} Y$ ← response

We saw that $\omega = \frac{2}{N}\sum_{i=1}^{N} \mathrm{Cov}(\hat{y}_i, y_i) = 2\frac{d}{N}\cdot\sigma_\varepsilon^2$ in other words

$$\sum_{i=1}^{N} \mathrm{Cov}(\hat{y}_i, y_i) = \mathrm{tr}(H)\cdot\sigma_\varepsilon^2$$

which leads to $\underbrace{\mathrm{tr}(H)}_{d} = \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{N}\mathrm{Cov}(\hat{y}_i, y_i)$

motivating a new definition of degrees of freedom as a generalization of the number of parameters in a model

For a linear smoother method $\hat{Y} = S\,Y$, the effective number of parameters is $df(S) = \mathrm{trace}(S)$ ← Ex 7.6 on kNN + 7.5 on S general

In Mallows C_p we just replace d with $\mathrm{trace}(S)$.

And the more general def of $df(\hat{y})$ is

$$df(\hat{y}) = \frac{\sum_{i=1}^{N} \mathrm{Cov}(\hat{y}_i, y_i)}{\sigma_\varepsilon^2}$$

more in Part 2 for other $\hat{y}$ (ridge, lasso)

Exercises

Expected training and test MSE for linear regression

Do exercise 2.9.

Important take home message: We have proven (for MLR) that the expected test MSE is always at least as large as the expected training MSE.

Establish the average optimism in the training error

$$\omega = \frac{2}{N} \sum_{i=1}^{N} \mathrm{Cov}(\hat{y}_i, y_i)$$

Exercise 7.4

Plan

1) Look at EPE(x_0) (now called Err(x_0)) and how model complexity can be broken down into irreducible error, squared bias and variance (should be known from before)
2) Study EPE (Err) unconditional and conditional on the training set
3) Study optimism of the training error rate, and how in-sample error may shed light
4) Cross-validation and .632 bootstrap estimates of EPE + maybe start on missing data analysis!
5) How will we build on this in the rest of the course?

We finished 1) and 2) in L2, now we continue!