MA8701 Advanced methods in statistical inference and learning

L3-4: Model selection and assessment

Mette Langaas

1/8/23

Course homepage:
https://wiki.math.ntnu.no/ma8701/2023v/start

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
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
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

(again - focus is on regression)

First, nothing new, but new notation (X^0, Y^0) to specify that a new test observation is drawn from the joint distribution F (both over new X and new Y):

$$\text{Err}_T = \text{E}_{X^0,Y^0}[L(Y^0, \hat{f}(X^0)) \mid T]$$

and then the averaging over the training set (both Xs and Ys in the training set):

$$\text{Err} = \text{E}_T\text{E}_{X^0,Y^0}[L(Y^0, \hat{f}(X^0)) \mid T]$$

This is also called *extra-sample error* (in contrast to what we now will define to be in-sample).

We saw before - from the ESL Figure 7.1, the training error $\overline{\text{err}}$ is (in general) less than (or equal to) the true test error, so not a good estimator for the test error.

$$\overline{\text{err}} = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$

[In Exercise 2.9 we prove that the expected training error is smaller or equal the expected error of a testset - for MLR. Important to work on this exercise!]

Part of this is due to where the X values are "placed". The test input vectors need not be "in the same positions" as in the training X values (when the mean is taken over the full distribution of X).

To eliminate this "confusing fact", calculations can be made be assuming the X-values in the training data are kept fixed - and this is called the *in-sample error*. (We did the same in TMA4267 using the Fahrmeir et al book, Chapter 3.4.)

In-sample error

$$\mathsf{Err}_{\mathsf{in}} = \frac{1}{N}\sum_{i=1}^{N} \mathsf{E}_{Y^0}[L(Y_i^0, \hat{f}(x_i)) \mid T]$$

Observe that we now take the expected value over distribution of the response - but that the (new) responses are found at the original training points. The training predictor positions x_i, $i = 1, \ldots, N$ are fixed. In addition the responses in the training data are also kept fixed, so the only random quantity here is the new responses at the fixed predictors.

Optimism

Optimism is defined as the difference between the in-sample error and the training error:

$$\text{op} = \text{Err}_{\text{in}} - \overline{\text{err}}$$

Average optimism

is defined as the expected value of the optimism, where the expectation is taken over the distribution of the training responses - denoted $\mathbf{y}$ (training predictors still kept fixed):

$$\omega = \text{E}_{\mathbf{y}}(op) = \text{E}_{\mathbf{y}}(\text{Err}_{\text{in}}) - \text{E}_{\mathbf{y}}(\overline{\text{err}})$$

Observe that if we write T then the expectation is taken over the distribution of both the predictors and responses in the training set, and we here write $\mathbf{y}$ for taking the distribution only over the reponses in the training set (not the predictors in the training set). So: we will focus on "modelling" ω, "instead of" Err.

Covariance result

For squared error (see ESL Exercise 7.4), 0-1 loss, and "other loss functions" it can be shown

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

Interpretation:

- how much the training error *underestimates* the true error depends on how strongly the observed response y_i affects its own prediction $\hat{y}_i$.
- the *harder* we fit the data the greater the covariance - which increases the expected (averaged) optimism.

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!

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? We look at a generalization in ESL exercise 7.5.
Observe that the optimism increases with d and decreases with N.
Comment: versions of the formula hold approximately for other error models than linear with squared loss (ESL mention binary data and entropy loss), but not in general for 0-1 loss (page 231, bottom, with reference to Efron 1986 - consult the ESL book).

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

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

- $P_{\hat{\theta}}(Y)$: family of density for Y where the true density is included
- $\widehat{\theta}$: 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.

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

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

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

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.

The effective number of parameters

(ESL 7.6)

The number of parameters d can be generalized into an *effective number of parameters*. We will look at linear fitting method:

$$\hat{\mathbf{y}} = \mathbf{S}y$$

where $\mathbf{S}$ as a $n \times n$ matrix depending on covariates x_i but not responses y_i.

- MLR $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$
- cubic smoothing splines
- ridge regression

The effective number of parameters is

$$\mathrm{df}(\mathbf{S}) = \mathrm{trace}(\mathbf{S})$$

Remember that the trace of a square matrix is the sum of the diagonal elements, and trace is often denoted tr.

Additive error model and squared loss: $Y = f(X) + \varepsilon$ with $\text{Var}(\varepsilon) = \sigma_\varepsilon^2$ then

$$\sum_{i=1}^{N} \text{Cov}(\hat{y}_i, y_i) = \text{trace}(\mathbf{S})\sigma_\varepsilon^2$$

leading to a generalization

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

See exercise 7.5 to prove this.

We return to this formula when we look at neural networks with quadratic penalization (weigth decay, ridge regularization) in Part 2.

Cross-validation (CV)

(ESL Ch 7.10, 7.12 - most should be known from TMA4268)

The aim is to estimate Err_T, but from simulation analyses (ESL Ch 7.12) it turns out that cross-validation estimates Err "the best".

The starting point for the method is that we only have one training set - and try to use that for either model selection or model assessment (not both).

What to do when both is needed, is not covered in this chapter. Nested cross-validations aka two-layers of cross-validation is one possibility. Another is to set aside data for a test set for model assessment, but use the training set in cross-validation for model selection.

Formal set-up for model assessment

- The allocation of observation $\{1, \ldots, N\}$ to folds $\{1, \ldots, K\}$ is done using an indexing function $\kappa : \{1, \ldots, N\} \to \{1, \ldots, K\}$, that for each observation allocate the observation to one of K folds.
- Further, $\hat{f}^{-k}(x)$ is the fitted function, computed on the observations except the kth fold (the observations from the kth fold is removed).
- The CV estimate of the expected prediction error $\mathsf{Err} = \mathsf{Err} = \mathsf{E}_T \mathsf{E}_{X^0, Y^0}[L(Y^0, \hat{f}(X^0)) \mid T]$ is then

$$\mathsf{CV}(\hat{f}) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}^{-k(i)}(x_i))$$

Formal set-up for model selection

- The indexing function κ is unchanged, and for the fitting function we add a tuning parameter α: $f(x, \alpha)$ such that $\hat{f}^{-k}(x, \alpha)$ is the fitted function using tuning parameter α, with the kth fold removed from the model fitting.
- The expected prediction error is estimated by

$$\text{CV}(\hat{f}, \alpha) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}^{-k(i)}(x_i, \alpha))$$

- We find the best tuning parameter $\hat{\alpha}$ that minimize the $\text{CV}(\hat{f}, \alpha)$. Alternatively the *one-standard error rule* can be used: choose the most parsimonious ("smallest") model whose error is no more than one standard error above the error of the best model.
- This best chosen model is then fit to all the data. (ESL page 242).

Choice of K

- Popular choices are 5 and 10 based on observations in simulation studies- and arguments similar to a bias-variance trace off.
- $K = N$ is called *leave-one-out* cross-validation LOOCV, and gives the lowest bias for estimating the Err.

Generalized cross-validation (GCV)

For LOOCV with squared loss and linear fitting. Remember

$$\hat{\mathbf{y}} = \mathbf{S}y$$

For many fitting methods (including MLR)

$$\frac{1}{N}\sum_{i=1}^{N}[y_i - \hat{f}^{-i}(x_i)]^2 = \frac{1}{N}\sum_{i=1}^{N}[\frac{y_i - \hat{f}(x_i)}{1 - S_{ii}}]^2$$

where S_{ii} is the ith diagonal element of $\mathbf{S}$. This leads to the GCV approximation:

$$\text{GCV}(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N}[\frac{y_i - \hat{f}(x_i)}{1 - \text{tr}(\mathbf{S})/N}]^2$$

where we recognise the effective number of parameters trace$(\mathbf{S})$. In some settings the trace$(\mathbf{S})$ is computed more easily than the individual elements S_{ii}.

The wrong and the right way to do cross-validation

In short: make sure that all part of the model fit process is "inside" the CV.

See learning material from TMA4268: Module 5: Resampling, and I also recommend to work on Problem 3 with solutions

Bootstrap methods

(ESL Ch 7.11 - bootstrapping is known from TMA4268 and TMA4300, but not the special case of estimating Err). Bootstrap in TMA4268: Module 5

Notation: $\mathbf{Z} = (z_1, \ldots, z_N)$ is the training set with $z_i = (x_i, y_i)$.

Aim: Of interest is some quantity calculated from the data $\mathbf{Z}$, denoted $S(\mathbf{Z})$. We will have focus on the expected prediction error.

Resampling: We draw with replacement from $\mathbf{Z}$ a total of N observastions into $\mathbf{Z}^{*b}$. We repeat this B times.

Estimator for expected predicted error Err:

$$\widehat{\mathsf{Err}}_{\mathsf{boot}} = \frac{1}{B}\frac{1}{N}\sum_{b=1}^{B}\sum_{i=1}^{N} L(y_i, \hat{f}^{*b}(x_i))$$

However - $\widehat{\text{Err}}_{\text{boot}}$ is not a good estimator: bootstrap datasets are acting as training data and the original data as a test sample - and the two samples have observations in common.

This overlap can make predictions too good. Remeber, in CV we have no overlap.

Q: What is the probability that observation i is included in bootstrap sample b?

The problem is given in TMA4268 Module 5 as Problem 1 with (handwritten) solutions.

The answer is $1 - (1 - \frac{1}{N})^N \approx 1 - e^{-1} = 0.632$.

Why is this relevant?

What if we try to change the bootstrap Err estimator - so that we for each observation i only keep predictions from bootstrap samples this observation is not present? Then we would mimick the CV-estimator.

The *leave-one-out* bootstrap estimate:

$$\widehat{\mathsf{Err}}^{(1)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} L(y_i, \hat{f}^{*b}(x_i))$$

where C^{-i} are the indices in the bootstrap sample b that do not contain observation i, and $|C^{-i}|$ is the number of samples. (B must be large enough that we do not get any C^{-i}s that are empty, or leave out these zero sets in the formula.)

Comment: this is also called out-of-bootstrap, and is closely connected to the popular out-of-bag estimate for random forests.

There is an addition fix to make the estimate even better.

Since the average number of distinct observations in each bootstrap sample is approximately $0.632N$ - and the bootstrap sample behaves like a training set - this gives a socalled training-set-size bias (similar to C with $K = 2$), meaning that the leave-one-out bootstrap estimator will be *biased upwards*. This can be fixed by weighing together the leave-one-out boostrap estimator with the training error.

The ".632" estimator:

$$\widehat{\mathsf{Err}}^{(.632)} = 0.368\overline{\mathsf{err}} + 0.632\widehat{\mathsf{Err}}^{(1)}$$

According to ESL (page 251): the derivation of the .632 estimator is complex, and the estimator is expected to work well in situation where the data is not overfitted, but may break down in overfit situations.

According to CASI (page 323) the .632 rule is less variable than the leave-one-out CV.

Example of this on page 251-252: two equal size classes where predictors independent of class, classification with 1NN gives $\overline{\text{err}} = 0$, $\widehat{\text{Err}}^{(1)} = 0.5$ and thus $\widehat{\text{Err}}^{(.632)} = 0.632 \cdot 0.5 = 0.316$, where here the true error rate is 0.5.

There is an improved version of the estimator - taking into account the amount of overfitting, leading to an adjustment to the weight $w = 0.632$ (and $1 - w = 0.368$) dependent on a socalled *no-information error rate*=γ=the error rate of the prediction rule when predictors and class labels are independent.

$$\hat{\gamma} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{i'=1}^{N} L(y_i, \hat{f}(x_{i'}))$$

Further the *relative overfitting rate* is defined to be

$$\hat{R} = \frac{\widehat{\mathsf{Err}}^{(1)} - \overline{\mathsf{err}}}{\hat{\gamma} - \overline{\mathsf{err}}}$$

Finally, the ".632+"-estimator is

$$\widehat{\mathsf{Err}}^{(.632+)} = (1-\hat{w})\overline{\mathsf{err}} + \hat{w}\widehat{\mathsf{Err}}^{(1)}$$

where $\hat{w} = \frac{0.632}{1-0.368\hat{R}}$.

For details on this approach consult ESL page 252-253.

Conclusions: Model selection and assessment and future use

Group discussion: Construct a "mind map"/"overview sheet"/"concept map" for the "Model assessement and selection" topics, and write down important take home messages!

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

Exercise 7.4

Relate the covariance to the trace of a linear smoother

Exercise 7.5

Need to know about covariance and variance of linear combinations. The reading list in TMA4267 included Härdle and Simar (2015): Applied Multivariate Statistical Analysis (fourth edition) - ebook from Springer available at NTNU. Alternatively [classnotes from TMA4267 (page 58-59)] (https://www.math.ntnu.no/emner/TMA4267/2017v/TMA4267V2017Pa

Something with programming and data?

TO BE ADDED!

Solutions to exercises

Please try yourself first, or take a small peek - and try some more - before fully reading the solutions. Report errors or improvements to Mette.Langaas@ntnu.no. (The solutions given here are very similar to the UiO STK-IN4300 solutions, see link under References.)

- 2.9
- 7.4
- 7.5
- Programming and data example?

Reference links

- ESL official errata: and choose “Errata” in the left menu
- ESL solutions to exercises
- ESL solutions from UiO
- CASI Computer Age Statistical Inference, Efron and Hastie (2017). Chapter 12: Cross-Validation and C_p Estimates of Prediction Error
- Burnham and Andersen (2002): Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer. Chapter 7: Statistical Theory and Numerical Results

Bibliography

Efron, Bradley, and Trevor Hastie. 2016. *Computer Age Statistical Inference - Algorithms, Evidence, and Data Science*. Cambridge University Press. https://hastie.su.domains/CASI/.