MA8701 Advanced methods in statistical inference and learning

L2: Classification and statistical decision theory, model selection and assessment

Mette Langaas

1/10/23

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

Plan

Continue with the decision theoretic framework from L1, but now for classification. Bias-variance trade-off.

- Classification - should not be new (ESL Ch 4.1-4.5, except 4.4.4)
- Statistical decision theoretic framework for classification (ESL 2.4)
- and the bias-variance trade-off (ESL 2.9 and 7.2-7.3)

Statistical decision theoretic framework

(ESL Ch 2.4)

is a mathematical framework for developing models f - and assessing optimality.

Last time, regression:

- $X \in \Re^p$
- $Y \in \Re$
- $P(X, Y)$ joint distribution of covariates and respons

Aim: find a function $f(X)$ for predicting Y from some inputs X.

Ingredients: Loss function $L(Y, f(X))$ - for *penalizing errors in the prediction*.

Criterion for choosing f: Expected prediction error (EPE)

$$\mathsf{EPE}(f) = \mathsf{E}_{X,Y}[L(Y, f(X))] = \int_{x,y} L(y, f(x))p(x, y)dxdy$$

Choose f to minimize the $\mathsf{EPE}(f)$.

What changes do we need to do for classifiation?

Classification set-up

- $X \in \Re^p$
- $G \in G = \{1, \dots, K\}$
- $\hat{G}(X) \in G = \{1, \dots, K\}$ (why f for regression and $\hat{G}$ for classification? strange?)
- $L(G, \hat{G}(X))$ is a $K \times K$ matrix where $K = |G|$, with elements l_{jk} giving the price to pay to misclassify an observation with true class g_j to class g_k.
- Elements on the diagonal of L is 0, and off-diagonal elements are often 1.

We would like to find $\hat{G}$ to minimize the EPE:

$$\begin{aligned} \mathsf{EPE} &= \mathsf{E}_{G,X}[L(G,\hat{G}(X))] = \mathsf{E}_X \mathsf{E}_{G|X}[L(G,\hat{G}(X))] \\ &= \mathsf{E}_X \{\sum_{k=1}^{K} L(g_k,\hat{G}(X)) P(G = g_k \mid X = x)\} \end{aligned}$$

Also here it is sufficient to minimize the loss for each value of x (pointwise)

$$\hat{G} = \text{argmin}_{g \in G} \sum_{k=1}^{K} L(g_k, \hat{G}(X)) P(G = g_k \mid X = x)$$

In the special case of 0-1 loss (off-diagonal elements in L equal to 1) then all k except the correct class gives loss 1 with probability $P(G = g_k \mid X = x)$. Summing over the wrong classes gives the same as taking 1 minus the conditional probability of the correct class g.

$$\hat{G} = \mathsf{argmin}_{g \in G}[1 - P(G = g \mid X = x)]$$

$$= \mathsf{argmax}_{g \in G} P(G = g \mid X = x)$$

The *Bayes classifier* classifies to the most probable class using the conditional distribution $P(G \mid X)$. The class boundaries are the *Bayes decision boundaries* and the error rate is the *Bayes rate*.

Group discussion

1) What do we know about classification? (TMA4268 and TMA4315 mainly, or ESL ch 4.1-4.5, except 4.4.4)

Some possible variants:

- What is the difference between discrimination and classification?
- What are the sampling vs diagnostic paradigm? Which paradigm for kNN and LDA?
- Parametric vs non-parametric methods?

2) Logistic regression is by many seen as the "most important method in machine learning". What do we remember about logistic regression?

Resources

(mostly what we learned in TMA4267, or ESL ch 4.1-4.5, except 4.4.4)

- From TMA4268: Overview and in particular Module 4: Classification and Module 2: Statistical learning
- From TMA4315: Overview and in particular Module 3: Binary regression and for more than two classes: Module 6: Categorical regression.

Model assessment and selection

(ESL Ch 7.1-7.6,7.10-7.12)

We use a training set to estimate $\hat{f}$.

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

If we are in a *data rich situation* we “just” divide our data into three parts, and use

- one for training
- one for validation (model selection)
- one for testing (model assessment)

A typical split might be 50-60% training and 20-25% validation and test, but this depends on the complexity of the model to be fitted and the signal-to-noise ratio in the data.

The focus in Ch 7 of ESL is to present methods to be used in the situations where we *do not have enough data* to rely on the training-validation-testing split.

And, even if we have enough data - what we now will learn will give us insight into much used methods for model assessment and model selection!

Plan

1) Look at EPE(x_0) (now called Err(x_0) after we have estimated f) 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 on methods for model selection (like AIC, Mallows Cp)
4) Cross-validation and .632 bootstrap estimates of EPE
5) How will we build on this in Parts 2-4?

The bias-variance trade-off

(ESL p 26 and 7.3)

Assume additive error model:

$$Y = f(X) + \varepsilon$$

where $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2_\varepsilon$.

For the bias-variance decomposition we only consider the squared loss. Why?

In Ch 7 we use the notation Err instead of EPE (expected prediction error) that we used in Ch 2, and now we have estimated f by $\hat{f}$.

Let $\mathrm{Err}(x_0)$ be the expected prediction error of a regression fit $\hat{f}(X)$ at a (new) input value $X = x_0$. As in Ch 2 the expected value is over (X, Y) for Err, and we may look at

$$\mathrm{Err} = E_{x_0} \mathrm{Err}(x_0)$$

How can we partition $\mathrm{Err}(x_0)$ into different sources?

$$\mathsf{Err}(x_0) = \mathsf{E}[(Y - \hat{f}(x_0))^2 \mid X = x_0] = \sigma_\varepsilon^2 + \mathsf{Var}[\hat{f}(x_0)] + [\mathsf{Bias}(\hat{f}(x_0))]^2$$

- First term: irreducible error, $\mathsf{Var}(\varepsilon) = \sigma^2$ and is always present unless we have measurements without error. This term cannot be reduced regardless how well our statistical model fits the data.
- Second term: variance of the prediction at x_0 or the expected deviation around the mean at x_0. If the variance is high, there is large uncertainty associated with the prediction.
- Third term: squared bias. The bias gives an estimate of how much the prediction differs from the true mean. If the bias is low the model gives a prediction which is close to the true value.

Group activity

- Remind yourself on how this derivation was done and the meaning of each term.
- What is the role of x_0 here?

Derivation

If you need to refresh your memory of the bias-variance trade-off, you might also look at the exam Problem 2 TMA4268 2018 exam with solutions

Also: TMA4268 and in particular Module 2

The following is a derivation:

$$
\begin{aligned}
\mathsf{Err}(x_0) &= \mathsf{E}[(Y - \hat{f}(x_0))^2 \mid X = x_0] \\
&= \mathsf{E}[Y^2 + \hat{f}(x_0)^2 - 2Y\hat{f}(x_0) \mid X = x_0] \\
&= \mathsf{E}[Y^2 \mid X = x_0] + \mathsf{E}[\hat{f}(x_0)^2 \mid X = x_0] - \mathsf{E}[2Y\hat{f}(x_0) \mid X = x_0] \\
&= \mathsf{Var}[Y \mid X = x_0] + \mathsf{E}[Y \mid X = x_0]^2 + \mathsf{Var}[\hat{f}(x_0) \mid X = x_0] + \mathsf{E}[\hat{f}(x_0 \\
&= \mathsf{Var}[Y \mid X = x_0] + f(x_0)^2 + \mathsf{Var}[\hat{f}(x_0) \mid X = x_0] + \mathsf{E}[\hat{f}(x_0) \mid X = \\
&= \mathsf{Var}[Y \mid X = x_0] + \mathsf{Var}[\hat{f}(x_0) \mid X = x_0] + (f(x_0) - \mathsf{E}[\hat{f}(x_0) \mid X = \\
&= \mathsf{Var}(\varepsilon \mid X = x_0) + \mathsf{Var}[\hat{f}(x_0) \mid X = x_0] + [\mathsf{Bias}(\hat{f}(x_0)) \mid X = x_0]^2
\end{aligned}
$$

(For some applications also the training Xs are fixed.) See the exercises below to study the results for kNN and OLS.

Expected prediction error

(ESL 7.2 and 7.4, and we are now back to a general loss function - but first have regression in mind)
If we now keep the training set fixed (we would do that in practice - since we often only have one training set):

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

as before the expected value is with respect to (X, Y), but the training set is fixed - so that this is the test set error is for this specific training set T.
Getting back to the unconditional version, we take expected value over ALL that is random - now including the training set

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

We want to estimate Err_T, but we will soon see that it turns out that most methods estimate Err.

Training error

(also referred to as apparent error)
For a regression problem: The training error is the average loss over the training sample:

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

Group discussion

Look at Figure 7.1 (with figure caption) on page 220 in the ESL book. The text reads that “100 simulated training sets of size 50” and that “lasso produced sequence of fits” (this means that we have different model complexities on the x-axis).

Explain what you see - in particular what are the red and blue lines and the bold lines. What can you conclude from the figure?

- Red lines=
- Bold red line=
- Blue lines=
- Bold blue line=

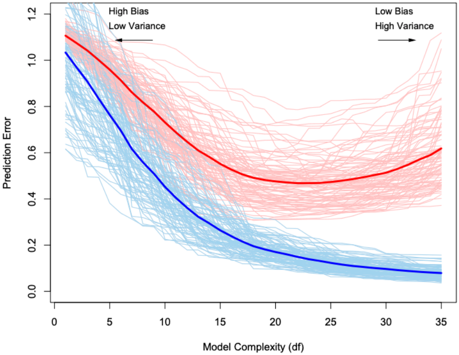

FIGURE 7.1. *Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error* $\overline{\text{err}}$, *while the light red curves show the conditional test error* $\text{Err}_{\mathcal{T}}$ *for* 100 *training sets of size* 50 *each, as the model complexity is increased. The solid curves show the expected test error* Err *and the expected training error* $\text{E}[\overline{\text{err}}]$.

Conclusion

(from Figure 7.1)
The training error $\overline{\mathrm{err}}$ is not a good estimate for the Err_T nor the Err.

Loss function and training error for classification

- $X \in \Re^p$
- $G \in G = \{1, \dots, K\}$
- $\hat{G}(X) \in G = \{1, \dots, K\}$

0-1 loss with $\hat{G}(X) = \text{argmax}_k \hat{p}_k(X)$

$$L(G, \hat{G}(X)) = I(G \neq \hat{G}(X))$$

-2-loglikelihood loss (why -2?):

$$L(G, \hat{p}(X)) = -2\text{log}\hat{p}_G(X)$$

Test error (only replace $\hat{f}$ with $\hat{G}$):

$$\mathsf{Err}_T = \mathsf{E}[L(Y, \hat{G}(X)) \mid T]$$

$$\mathsf{Err} = \mathsf{E}[\mathsf{E}[L(Y, \hat{G}(X)) \mid T]] = \mathsf{E}[\mathsf{Err}_T]$$

Training error (for 0-1 loss)

$$\overline{\mathsf{err}} = \frac{1}{N} \sum_{i=1}^{N} I(g_i \neq \hat{g}(x_i))$$

Training error (for -2loglikelihood loss)

$$\overline{\mathsf{err}} = -\frac{2}{N} \sum_{i=1}^{N} \mathsf{log} \widehat{p}_{g_i}(x_i)$$

Discussion and conclusions

- What are key take home messages from today´s teaching session?
- What do you plan to do before the next teaching session?
- Feedback on today´s teaching session?

Exercises

What are the most important results from the “Statistical decision theoretic framework”?

- What are results to remember for regression and for classification?
- How would you use these results?

Look into the derivation for the bias and variance decomposition

- for kNN in Equation 7.10 and
- OLS in Equation 7.11 on pages 222-223 of ESL.

Bayes classier, Bayes decision boundary and Bayes error rate

Solve TMA4268 exam problem 9 in 2019 at https://www.math.ntnu.no/emner/TMA4268/Exam/V2019e.pdf

Key results from logistic regression

a) What are the three components of a generalized linear model?

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.

Look into the derivation for the bias and variance

kNN in Equation 7.10:

https://waxworksmath.com/Authors/G_M/Hastie/WriteUp/Weatherwax

OLS

in Equation 7.11 on pages 222-223 of ESL: We solved this in 2019 in https://www.math.ntnu.no/emner/TMA4268/2019v/CompEx/Compulso
variance_trade-off, where we compared with the ridge (to be studied in Part 2) and also plotted functions.

Bayes classifier, Bayes decision boundary and Bayes error rate

https:
//www.math.ntnu.no/emner/TMA4268/Exam/e2019sol.html

Key results from logistic regression

a) What are the three components of a generalized linear model?

Reference links

- ESL official errata: and choose “Errata” in the left menu
- ESL solutions to exercises
- Jarle Tufto´s lecture notes in TMA4315 GLM, 168 pages

Bibliography

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