Introduction to Statistical Machine Learning for Bioinformaticians and Medical Informaticians

SoSe 2022

Tutor: Dana Petracek

1	2	$\sum$

Marina Dittschar & Clarissa Auckenthaler

Assignment 5

(Submitted 13.06.2022)

Task 1

Show that formula 5.6 in ISLR:

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_X Y}$$

indeed minimizes $Var(\alpha X + (1-\alpha)Y)$**.**

We know that for the general case that X and Y are correlated, the variance can also be written as [5]:

$$Var(X+Y) = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY} \tag{1}$$

Following this, we multiply out the given variance term [7] which results in:

$$Var(\alpha \cdot X + (1-\alpha)Y) = \alpha^2\sigma_X^2 + (1-\alpha)^2\sigma_Y^2 + 2\alpha(1-\alpha)\sigma_{XY} \tag{2}$$

We want to find the minimum of the variance dependent on α, so we take the first derivative in regards to α.

$$\begin{aligned} & Var(\alpha X + (1-\alpha)Y)' \\ &= 2\alpha\sigma_X^2 - 2(1-\alpha)\sigma_Y^2 + 2\sigma_X Y - 4\alpha\sigma_{XY} \\ &= 2\alpha\sigma_X^2 - 2\sigma_Y^2 + 2\alpha\sigma_Y^2 + 2\sigma_X Y - 4\alpha\sigma_{XY} \end{aligned} \tag{3}$$

To find minima as well as maxima, we need to set the first derivative to zero.

$$2\alpha\sigma_X^2 - 2\sigma_Y^2 + 2\alpha\sigma_Y^2 + 2\sigma_X Y - 4\alpha\sigma_{XY} = 0 \tag{4}$$

We divide the whole equation by 2:

$$\alpha\sigma_X^2 - \sigma_Y^2 + \alpha\sigma_Y^2 + \sigma_X Y - 2\alpha\sigma_{XY} = 0 \tag{5}$$

In Equation 5, we have coloured in all terms containing α in blue. We now put all the terms not containing α on the right by subtracting them and divide by all the terms containing α. This results in the following equation:

$$\alpha = \frac{\sigma_Y^2 - \sigma_X Y}{\sigma_X^2 + \sigma_Y^2 - \sigma_X Y} \tag{6}$$

Which is the same as the one that we set out to prove. It only remains to show that this value for α is always negative. To find out the sign of a given extreme point in a function, it needs to be determined whether the second derivative will be positive or negative at that point. To do this, we get the second derivative of the initial function [6]:

$$Var(\alpha X + (1-\alpha)Y)'' = 2\sigma_X^2 - 2\sigma_Y^2 - 4\sigma_{XY} = 2Var(X-Y) \tag{7}$$

We know that the variance can never be negative, and thus the given α minimizes the variance term. q.e.d.

Task 2

Selecting an overall optimal model in the subset selection methods requires to compare models of different complexity using model selection criteria. Here, your task is to study the model selection criteria and to explain definitions, similarities, and differences. Since metrics that were previously discussed in the lecture (Root Mean squared error ($RMSE$), R-squared (R^2), residual standard error (RSE) and Mean absolute error (MAE) are very sensitive to the number of variables, although they do not necessarily have a significant influence on the response, more robust metrics are needed to be able to compare models, since the different models can also have different complexity. In the following we compare the criteria Mallow's C_p, Akaike information criterion (AIC), Bayesian information criterion (BIC) and Adjusted R^2.

The definition for Mallows' C_p Criterion is:

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2) \tag{8}$$

In equation 8 we add $2d\hat{\sigma}^2$ to compensate the difference between test error and train error [2]. Given that $\hat{\sigma}^2$ is an unbiased estimate of σ^2 shows that C_p corresponds to an unbiased estimate of test MSE. The Mallows' C_p criterion needs to be minimized, which means minimizing the test error. In the end, the model is selected for which the smallest C_p value was calculated.

The definition for the Akaike information criterion (AIC) in the case of least squared models is as follows [2]:

$$AIC = \frac{1}{n}(RSS + 2d\hat{\sigma}^2) \tag{9}$$

In the case that maximum likelihood is not equivalent to least squares the equation is defined as [1]:

$$AIC = 2k - 2ln(\hat{L}) \tag{10}$$

with the number of independently adjusted parameters k and maximum likelihood $\hat{L}$. In the general case, AIC is defined for models fitted with a maximum likelihood function.

However, in the case that the model is fitted with Gaussian error distribution, maximum likelihood equals the least squared model (see equation 9). C_p and AIC are the same for known $\hat{\sigma}^2$. We select the best model according to the AIC by minimizing it.
Another model selection criterion is the Bayesian information criterion (BIC), which for the least squares model is defined as [2]:

$$BIC = \frac{1}{n}(RSS + log(n)d\hat{\sigma}^2) \tag{11}$$

And in the general case as:

$$BIC = kln(n) - 2ln(\hat{L}) \tag{12}$$

Minimizing the BIC is equivalent to maximizing the posterior model likelihood for a large collection of observations and can be used as a model selection criterion [8]. Similar to the AIC, the general formula (see equation 12) can be simplified if the maximum likelihood is equivalent to the least squares model (see equation 11). BIC is not as good for large sets of variables, because it generally leads to a stronger penalty of models with many parameters [2].

The adjusted R^2 is defined as [2]:

$$Adjusted \quad R^2 = 1 - \frac{\frac{RSS}{n-d-1}}{\frac{TSS}{n-1}} \tag{13}$$

with number of data points n and number of predictors d. Contrary to the unadjusted R^2, the *adjusted* R^2 does not necessarily decrease with growing number of samples [2]. The *adjusted* R^2 therefore penalizes a large number of parameters [2]. Maximizing the adjusted R2, or equivalently minimizing the MSE, is one way to determine a best model[3].

Similarities and Differences:

All of the criteria above can be used for model evaluation and selection. In addition to this trivial similarity, all of them are unbiased estimate of the model prediction error MSE [2]. For C_p, AIC and BIC the problem is of minimization, which means a small value indicates a model with a low test error [8]. In contrast to this, the Adjusted R^2 is a maximization problem, which means a large value indicates the best model (with small test error) [2]. AIC, C_p and BIC do look very similar in their definition (only for AIC and BIC when the maximum likelihood is equivalent to least squares, shown in equations 11, 9 and 8). One problem that occurs by using BIC for a large sample size is, that the model complexity is more penalized by BIC compared to the other discussed selection criteria [4], which means using C_p, adjusted R^2 or AIC is better for a large sample size than using BIC for the model evaluation and selection. All of the above mentioned metrics are easy to compute and use [2].

References

[1] H. Akaike. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6):716–723, 1974.

[2] G. James, D. Witten, T. Hastie, and R. Tibshirani. An Introduction to Statistical Learning: with Applications in R. Springer Texts in Statistics. Springer US, 2021.

[3] Purdue University. Model building: Selection criteria: Stat 512 , spring 2011: Background reading knnl: Chapter 9.

[4] Marco Taboga. Model selection criteria, lectures on probability theory and mathematical statistics. kindle direct publishing. online appendix. https://www.statlect.com/fundamentals-of-statistics/model-selection-criteria, 2021.

[5] Marco Taboga. 4.7: Variance sum law ii - correlated variables. https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_Introductory_Statistics_(Lane)/04%3A_Describing_Bivariate_Data/4.07%3A_Variance_Sum_Law_II_-_Correlated_Variables, 2022.

[6] Wikipedia. Extremwert. https://de.wikipedia.org/wiki/Extremwert#Beispiele, 2022.

[7] Wikipedia. Variance. https://en.wikipedia.org/wiki/Variance#Basic_properties, 2022.

[8] Ernst Wit, Edwin den van Heuvel, and Jan-Willem Romeijn. ‘all models are wrong...’: an introduction to model uncertainty. Statistica Neerlandica, 66(3):217–236, 2012.