Trial exam STK2100 (solutions) - spring 2022

Geir Storvik

Spring 2021

Exercise 1

(a) It is difficult to evaluate whether the fit is good based on the log-likelihood due to that there is no fixed scale on that. Further, the confusion matrix is evaluated on the same data as the fitting and might be too optimistic.

However, the regression table gives the impression that there are too many variables included indicating that some kind of variable selection should be performed.

(b) The maximum likelihood value will always be larger for a model with more variables included due to that we maximise over a larger space.

We can use the AIC kriterion to compare the models. This gives

$$\begin{aligned}\text{AIC}_1 &= -2*(-97.83)+2*(10) = 215.66\\ \text{AIC}_2 &= -2*(-98.71)+2*(6) = 209.42\end{aligned}$$

The second model has almost as good likelihood value with less parameters, which results in that the AIC value for the new model is lower and the one to prefer.

If we look at the confusion matrix for the new model, it has one more error, but given that this is on the training data this can be due to overfitting on the first (larger) model.

(c) The P-values are values used for testing the corresponding hypothesis $H_{0j} : \beta_j = 0$ *conditioned on the other parts of the model is included.* We see that while the first model has many large P-values and thereby do not give significant reasons for rejecting H_{0j} for many j's, the second model has all small P-values indicating that all variables now are important.

A possible reason for the P-values to change in the second model is that the corresponding covariates are correlated with some of the variables that have been removed.

(d) We have that each Y_i is binomial distributed with one trial. The probability for the response to be equal to one vary from one observation to the next. This is then marked by an index i on p_i. By in addition assume independence between the responses, we obtain products of the form $p_i^{y_i}(1-p_i)^{1-y_i}$.

Since we for classification trees assume that the probabilities are equal within each region, we obtain that $p_i = c_m$ for $\boldsymbol{x}_i \in R_m$.

(e) We have 5 leaves which give 5 c_m parameters. In addition we have 4 splits. Each split has two parameters, one specifying which variable that the split depend on, the other which value the split is made on. In total we then get $5 + 2 * 4 = 13$ parameters.

This gives an AIC value

$$AIC = -2 * (-90.21) + 2 * 13 = 206.43$$

that is a bit better than what we obtained earlier.

Note that for classification trees, usually the complexity is rather measured by the number of leaves, calculating

$$AIC = -2 * (-90.21) + 2 * 5 = 190.43$$

which gives an even better measure of the model.

(f) We now see that the error rates are larger than previously. This is due to that we did not take into account overfitting earlier.

Now we see that logistic regression is doing better than classification tree with respect to error rate. We therefore will prefer logistic regression in this case.

(g) With Bagging we make many classifiers based on bootstrap samples from the original data. For each classifier we get a classification and we can then combine the classifications by counting which class that appears most often.

When we make bootstrap samples, some observations will not be part of the training set. These can be used for validation/testing. When we then look over all bootstrap samples, some (most) observations will be in a test set several times. We can then do a final classification by looking at those predicted classes that appears most times.

We see that for logistic regression we get no change in the confusion matrix while for the classifiation tree we get some improvements. This is due to that logistic regression do not have that much variability and then it does not help much in doing bagging. It is the opposite case for classification trees.

Exercise 2

(a) This is called Ridge regression. The main difference is that we introduce a penalty on the β-coefficients, in particular shrinking them to zero. This reduces the variance, at a potential loss in increased bias.

(*b*) For simplicity, assume $\sum_{i=1}^{N} x_{ij} = 0$. Defining $L_\lambda(\boldsymbol{\beta})$, we have

$$\begin{aligned}\frac{\partial}{\partial \beta_0} L_\lambda(\boldsymbol{\beta}) =& -2\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}) \\ =& -2\sum_{i=1}^{N} y_i + 2\sum_{i=1}^{N}\beta_0 + 2\sum_{i=1}^{N}\sum_{j=1}^{p}\beta_j x_{ij} \\ =& -2\sum_{i=1}^{N} y_i + 2\sum_{i=1}^{N}\beta_0\end{aligned}$$

and putting this to zero gives $\hat{\beta}_0 = \bar{y}$. Further, for $k = 1, 2, .., p$,

$$\begin{aligned}\frac{\partial}{\partial \beta_k} L_\lambda(\boldsymbol{\beta}) =& -2\sum_{i=1}^{N}(y_i - \bar{y} - \sum_{j=1}^{p}\beta_j x_{ij})x_{ik} + 2\lambda\beta_k \\ =& -2\sum_{i=1}^{N} y_i x_{ik} + 2\sum_{i=1}^{N}\bar{y}x_{ik} + 2\sum_{i=1}^{N}\sum_{j=1}^{p}\beta_j x_{ij}x_{ik} + 2\lambda\beta_k \\ =& -2\sum_{i=1}^{N} y_i x_{ik} + 2\sum_{i=1}^{N}\sum_{j=1}^{p}\beta_j x_{ij}x_{ik} + 2\lambda\beta_k\end{aligned}$$

which, when put to zero, gives the equation system

$$\sum_{j=1}^{p}\beta_j\sum_{i=1}^{N} x_{ij}x_{ik} + \lambda\beta_k = \sum_{i=1}^{N} y_i x_{ik}, \qquad\qquad k = 1, ..., p$$

or in vector/matrix form (with now $\boldsymbol{X}$ *not* including a first column with 1's)

$$[\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}]\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{Y}$$

and thereby $\widehat{\boldsymbol{\beta}} = [\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}]^{-1}\boldsymbol{X}^T\boldsymbol{Y}$.

Alternatively, one could directly write (after inserting $\hat{\beta}_0 = \bar{y}$)

$$\begin{aligned}L_\lambda(\boldsymbol{\beta}) =& (\boldsymbol{Y} - \bar{y}\mathbf{1} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \bar{y}\mathbf{1} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta} \\ =& (\boldsymbol{Y} - \bar{y}\mathbf{1})^T(\boldsymbol{Y} - \bar{y}\mathbf{1}) - 2((\boldsymbol{Y} - \bar{y}\mathbf{1})^T\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta}\end{aligned}$$

and take the derivative of $\boldsymbol{\beta}$ directly,

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\beta}} L_\lambda(\boldsymbol{\beta}) =& -2(\boldsymbol{Y} - \bar{y}\mathbf{1})^T\boldsymbol{X} + 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{\beta}^T \\ =& -2\boldsymbol{Y}^T\boldsymbol{X} + 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X} + 2\lambda\boldsymbol{\beta}^T\end{aligned}$$

since $\mathbf{1}^T\boldsymbol{X} = \mathbf{0}$.

(*c*) This is the Lasso method. Although also in this case the β_j's are shrink to zero, the form of penalization in this case results in that some parameter estimates can be exactly zero. Due to this we have that the left hand plot correspond to Lasso. Further, the least squares estimates are obtained by looking at the rightmost values.

(*d*) The λ value should be based on some evaluation set, most efficiently performed by cross-validation.

Exercise 3

(*a*) The increase in number of light vehicles results in an increase in the response as well which is reasonable. The same is true for heavy vehicles.

For average velocity, we see that the influence from this variable only appears when the speed is above 40, again quite reasonable.

For temperature we see that when the temperature is below 0, it does not really matter how low it is, between 0 and 15 there is a high influence on temperature while it is smaller for temperatures above 15.

For rainfall last four hours, there is a negative influence up to a certain amount of rain, then it flattens out.

The rainfall last week seems to be less important, but with a somewhat similar structure as for the last four hours.

(*b*) We see that the speed gives an increase of about 0.3 per 10 kilometer, indicating that we get a reduction of approximately 0.6. Note however that we then are extrapolating outside the region of fitted curves which might be somewhat dangerous.

Exercise 4

(*a*) When training/learning a model (estimating parameters) we do so by making it fit as good as possible to the given dataset. If we use the same data for both training and evaluation, we will get a too optimistic measure on how well the model is doing on predicting new data. We therefore need a separate evaluation set to measure the real performance.

In many cases we compare many different models/methods and select the one that performs the best on the evaluation set. This is a reasonable approach for choice of model, but by choosing the minimum of a set of random variables (here the performance measure) we end again up with a measure that is too optimistic (too low). In order to take the model choice into account we therefor also need a separate test set for the final evaluation.

(*b*) A weakness when dividing data into training/validation is that we reduce the amount of data that we can use for training.

Cross-validation is a smart way of utilizing the given data efficiently. The procedure is to divide the data into K groups, train og $K-1$ groups and evaluate on the last

one. This can be circulated, in which case we always train on a fraction $(K-1)/K$ of the data while evaluation actually is performed on all the data.

If we consider several models, cross-validation can be used for choosing a model. In that case we again need a separate test set for evaluating the final model.

If we only have one model to consider, cross-validation will give an (almost) unbiased estimate of the prediction error.

Exercise 5

(a) We have a potential discontinuity in the point $x = c$. Enforcing continuity imply that

$$\beta_{0,1} + \beta_{1,1}c + \beta_{2,1}c^2 = \beta_{0,2} + \beta_{1,2}c + \beta_{2,2}c^2$$

while continuous derivatives imply

$$\beta_{1,1} + 2\beta_{2,1}c = \beta_{1,2} + 2\beta_{2,2}c.$$

We started with 6 parameters, but with 2 constraints, we end up with 4 *free* parameters.

(b) Within the interval $(-\infty, c)$ we have

$$g(x) = \theta_0 + \theta_1 x + \theta_2 x^2$$

while within the interval (c, ∞) we have

$$g(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 (x-c)^2$$

which both are quadratic functions.

Further we have that $g(c) = \theta_0 + \theta_1 x + \theta_2 c^2$ in both cases showing that g is continuous.

Within the interval $(-\infty, c)$ we have

$$g'(x) = \theta_1 + 2\theta_2 x$$

while within the interval (c, ∞)

$$g'(x) = \theta_1 + 2\theta_2 x + 2\theta_3 (x-c)$$

showing that $g'(c) = \theta_1 + 2\theta_2 c$ in both cases and thereby also has continuous derivatives.

(c) In the interval $(-\infty, c)$ we must have

$$\beta_{0,1} + \beta_{1,1}x + \beta_{2,1}x^2 = \theta_0 + \theta_1 x + \theta_2 x^2$$

which imples that $\theta_0 = \beta_{0,1}, \theta_1 = \beta_{1,1}, \theta_2 = \beta_{2,1}$.

In the intervall $[c, \infty)$ we must have

$$\begin{aligned}\beta_{0,2} + \beta_{1,2}x + \beta_{2,2}x^2 =& \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 (x-c)^2 \\ =& \beta_{0,1} + c^2\theta_3 + (\beta_{1,1} - 2c\theta_3)x + (\beta_{2,1} + \theta_3)x^2\end{aligned}$$

which imply 3 requirements on θ_3:

$$\begin{aligned}\theta_3 =& c^{-2}(\beta_{0,2} - \beta_{0,1}) \\ \theta_3 =& \tfrac{1}{2c}(\beta_{1,1} - \beta_{1,2}) \\ \theta_3 =& \beta_{2,2} - \beta_{2,1}\end{aligned}$$

However, if we use the constraints from (a) we have that

$$\begin{aligned}\tfrac{1}{2c}(\beta_{1,1} - \beta_{1,2}) =& \tfrac{1}{2c}(2\beta_{2,2}c - 2\beta_{2,1}c) = \beta_{2,2} - \beta_{2,1} \\ c^{-2}(\beta_{0,2} - \beta_{0,1}) =& c^{-2}[\beta_{1,1}c + \beta_{2,1}c^2 - (\beta_{1,2}c + \beta_{2,2}c^2)] \\ =& \beta_{2,1} - \beta_{2,2} + c^{-1}(\beta_{1,1} - \beta_{1,2}) \\ =& \beta_{2,1} - \beta_{2,2} + 2(\beta_{2,2} - \beta_{2,1}) = \beta_{2,2} - \beta_{2,1}\end{aligned}$$

which shows that we actually only have one requirement.

(d) We have potential non-continuities in the points $c_1, ..., c_{M-1}$. Continuity implies

$$\beta_{0,m} + \beta_{1,m}c_m + \beta_{2,m}c_m^2 = \beta_{0,m+1} + \beta_{1,m+1}c_m + \beta_{2,m+1}c_m^2$$

while continuous derivatives imply

$$\beta_{1,m} + 2\beta_{2,m}c_m = \beta_{1,m+1} + 2\beta_{2,m+1}c_m.$$

We start with $3M$ parameters, but with $2(M-1)$ constraints, we end up with $3M - 2(M-1) = M + 2$ free parameters.

(e) Since we can write the model as a linear combination of basis functions, the free parameters can be estimated be ordinary least squares.

Exercise 6

(a) We have that

$$\begin{aligned}\Pr(Y = 1|\boldsymbol{x}) =& \frac{\exp(\eta(\boldsymbol{x}))}{1 + \exp(\eta(\boldsymbol{x}))} \\ \eta(\boldsymbol{x}) =& \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \beta_4 x_2 + \beta_5 x_2^2 + \beta_6 x_2^3 + \beta_7 x_3 + \beta_8 x_3^2 + \beta_9 x_3^3\end{aligned}$$

Many of the P-values are very large, indicating that we should do some kind of model selection.

(*b*) AIC is defined by

$$AIC = -2 \log L(\hat{\boldsymbol{\theta}}) + 2p$$

where p is the number of parameters. In the stepwise AIC procedure one start with the full model and then remove the one variable that makes the most improvement on the AIC value. It stops when no improvements can be made. Alternatively one can go the opposite way.

In this case several of the components are removed and for the remaining components the P-values are much smaller. The latter is probably due to correlations between the covariates in the full model. For Number, only a linear term remains.

(*c*) For the BIC criterion we are instead using

$$BIC = -2 \log L(\hat{\boldsymbol{\theta}}) + \log(N)p$$

In this case $\log(N) = \log(81) = 4.39$ so that we have a higher penalty on complex models. This results in that BIC will favour more simple models compared to AIC, which we also see here.

(*d*) We saw that after model selection, Number was only included as a linear effect (or not iuncluded at all). From the second plot we see that we can fit a linear function within the confidence band, indicating that there is no siginficant non-linear structure of Number in this case. This linear function can actually be quite close to the zero-line, indicating that Number is not that important.

For the two other variables, the non-linear structure is more clear, corresponding to that we have higher order polynomials included in the models previously considered.

(*e*) We see that the first splits are based on Start and Age, indicating that these are the most important covariates, similar to what we saw when doing model selection. Looking at the boxplots, splitting by Start at the first split seems very reasonable. For the next splits it is more difficult to see due to that it then depends on what we have done in the first split, but the split by age also seems reasonable.

(*f*) Assume we are using AIC/BIC, we first need to derive the number of parameters.

For logistic we have 10 parameters.

For logistic selected with AIC we reduced it to 6 parameters and down to 4 based on BIC.

For GAM it is not quite clear how many parameters we have, but it should be at least as many as models consider in (b) and (c). Since we got worse log-likelihood value here, we would not prefer this model.

For the tree-based model, we have 6 leaves giving 6 parameters, 5 splits giving 10 parameters, in total 16 parameters. However, another (more common?) way to measure

complexity of trees is to only consider the number of leaves, that is using $p = 6$ in this case. We include this as an alternative

We then get

Method	Log-likelihood	AIC	BIC
Logistic (a)	-23.83	67.66	91.60
Logistic select AIC (b)	-24.77	61.54	75.91
Logistic select BIC (c)	-27.44	62.88	72.46
GAM (d)	-29.26	?	?
Tree (e)	-23.94	79.88	118.19
Tree (e), number of leaves	-23.94	59.88	74.25

We see that logistic select seems to be the best, but the Tree might also be good depending on how we measure complexity.

Exercise 7

(a) We then have to minimize (dropping for simplicity the dependence on $\boldsymbol{x}_0$)

$$g(\beta_0, \beta_1) = \sum_{i=1}^{N} K(x_i, x_0)(y_i - \beta_0 - \beta_1 x_i)^2$$

with respect to $\beta_0(x_0), \beta_1(x_0)$. This is a smooth function in the two parameters and we have

$$\begin{aligned}
\frac{\partial}{\partial \beta_0} g(\beta_0, \beta_1) &= -2 \sum_{i=1}^{N} K(x_i, x_0)(y_i - \beta_0 - \beta_1 x_i) \\
&= -2 \sum_{i=1}^{N} K(x_i, x_0) y_i + 2\beta_0 \sum_{i=1}^{N} K(x_i, x_0) + 2\beta_1 \sum_{i=1}^{N} K(x_i, x_0) x_i \\
\frac{\partial}{\partial \beta_1} g(\beta_0, \beta_1) &= -2 \sum_{i=1}^{N} K(x_i, x_0)(y_i - \beta_0 - \beta_1 x_i) x_i \\
&= -2 \sum_{i=1}^{N} K(x_i, x_0) y_i x_i + 2\beta_0 \sum_{i=1}^{N} K(x_i, x_0) x_i + 2\beta_1 \sum_{i=1}^{N} K(x_i, x_0) x_i^2
\end{aligned}$$

and putting these equations to zero we obtain

$$\begin{aligned}
\beta_0 \sum_{i=1}^{N} K(x_i, x_0) + \beta_1 \sum_{i=1}^{N} K(x_i, x_0) x_i &= \sum_{i=1}^{N} K(x_i, x_0) y_i \\
\beta_0 \sum_{i=1}^{N} K(x_i, x_0) x_i + \beta_1 \sum_{i=1}^{N} K(x_i, x_0) x_i^2 &= \sum_{i=1}^{N} K(x_i, x_0) y_i x_i
\end{aligned}$$

or

$$\begin{pmatrix} \sum_{i=1}^N K(x_i, x_0) & \sum_{i=1}^N K(x_i, x_0)x_i \\ \sum_{i=1}^N K(x_i, x_0)x_i & \sum_{i=1}^N K(x_i, x_0)x_i^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^N K(x_i, x_0)y_i \\ \sum_{i=1}^N K(x_i, x_0)y_i x_i \end{pmatrix}$$

which can be easily solved.

By using this method, we assume that (locally) $f(x)$ is linear. When this is not the case, we will obtain a bias.

When we predict y_i, we use $f(\boldsymbol{x}_i)$. We have that

$$f(\boldsymbol{x}_i) = \hat{\beta}_0(x_i) + \hat{\beta}_1(x_i)x_i$$

Now we see from the equation system above that the left hand side do not depend on the y's while the right hand side are linear functions of the y's. We then *do* get that also $f(\boldsymbol{x}_i)$ becomes a linear function of the y's.

For $d = 2$ we will obtain a similar equation system and the argument will then also be similar.

(b) When we have the structure $\widehat{y} = \boldsymbol{S}\boldsymbol{y}$, a way of measuring complexity (degrees of freedom) is to use Trace($\boldsymbol{S}$). By choosing these to be similar, we obtain approximately the same complexity.

From the plot it seems quite reasonable to assume a local linear structure for most of the x points. It is therefore also reasonable that a measure that gives the average (or sum) performance becomes best for $d = 1$. If one however was very interested in the point where the slope change, it might be that $d = 2$ would be better there!

(c) For the smoothing spline method, the penalization term is directly related to smoothness (or curvature) of the function so we enforce smoothness.

Given that we have chosen the complexities to be approximately similar, we should choose the one with best prediction performance which then is local regression with $d = 1$.

(d) This belongs to the class of additive models

The structure of the two plots are very similar, but the scales are quite different, indicating that when temperature is left out, the model tries to compensate for this by scaling up the wind part.

In both cases it is clear that increasing wind has a positive effect on the ozon level which is reasonable.

(e) One could start with putting $f_2^0(x_2) = 0$ and then estimate $f_1^1(x_1)$. Then one can define the residuals

$$r_i = y_i - f_1^1(x_1)$$

and use these residuals as response when estimating $f_2^2(x_2)$. We can then construct new residuals

$$r_i = y_i - f_2^2(x_1)$$

and estimate $f_1^3(x_1)$ by using these new residuals as responses, and so on.

This is called backfitting.

Exercise 8

(a) We have that

$$\begin{aligned}\boldsymbol{E} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} =& \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta} + \varepsilon) \\ =& [\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{\varepsilon}\end{aligned}$$

This directly gives $E[\boldsymbol{E}] = \boldsymbol{0}$.

We obtain the covariance matrix for $\boldsymbol{E}$ by

$$\begin{aligned}\text{Var}(\boldsymbol{E}) =& [\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\sigma^2\boldsymbol{I}[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]^T \\ =& [\boldsymbol{I} - 2\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T + \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\sigma^2 \\ =& [\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\sigma^2\end{aligned}$$

(b) We have

$$\begin{aligned}E[\text{RSS}] =& E[\boldsymbol{E}^T\boldsymbol{E}] \\ =& E[\text{trace}(\boldsymbol{E}\boldsymbol{E}^T)] \\ =& \text{trace}(E[\boldsymbol{E}\boldsymbol{E}^T]) \\ =& \text{trace}(\text{Var}(\boldsymbol{E})) \\ =& \text{trace}([\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\sigma^2) \\ =& \sigma^2\text{trace}(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T) \\ =& \sigma^2(\text{trace}(\boldsymbol{I}) - \text{trace}(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T) \\ =& \sigma^2(n - \text{trace}(\boldsymbol{I}_p)) = \sigma^2(n-p)\end{aligned}$$

(c) We have

$$\begin{aligned}\text{Cov}(\hat{\boldsymbol{y}}, \boldsymbol{E}) =& \text{Cov}(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, [\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{\varepsilon}) \\ =& \text{Cov}(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\varepsilon}, [\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{\varepsilon}) \\ =& \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\text{Cov}(\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon})[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T] \\ =& \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sigma^2\boldsymbol{I}[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T] \\ =& \sigma^2[\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T] \\ =& \sigma^2[\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T] = \boldsymbol{0}\end{aligned}$$

This means that the error we make is independent of the actual size of the prediction. Geometrically this means that $\hat{\boldsymbol{y}}$ is orthogonal to $\boldsymbol{E}$.

Exercise 9

(a) We assume ε has expectation 0. We further assume Y is a new observation that has not been used for construction of $\widehat{Y}$. We first have that

$$E[(Y-\hat{Y})^2] = E[E[(Y-\hat{Y})^2|\boldsymbol{x}]]$$

and it is enough to minimize $E[(Y-\hat{Y})^2]\boldsymbol{x}]$ for each $\boldsymbol{x}$ separately.

Further,

$$\begin{aligned}
&E[(Y-\hat{Y})^2|\boldsymbol{x}] \\
&=E[(Y-E[Y|\boldsymbol{x}]+E[Y|\boldsymbol{x}]-\widehat{Y})^2|\boldsymbol{x}] \\
&=E[(Y-E[Y|\boldsymbol{x}])^2|\boldsymbol{x}]+E[(E[Y|\boldsymbol{x}]-\widehat{Y})^2|\boldsymbol{x}]+2E[(Y-E[Y|\boldsymbol{x}])(E[Y|\boldsymbol{x}]-\widehat{Y})|\boldsymbol{x}] \\
&=\mathrm{Var}(Y|\boldsymbol{x})+E[E[Y|\boldsymbol{x}]-\widehat{Y}(\boldsymbol{x}))^2|\boldsymbol{x}] \\
&=\mathrm{Var}(\varepsilon)+E[f(\boldsymbol{x})-\widehat{Y}(\boldsymbol{x}))^2|\boldsymbol{x}]
\end{aligned}$$

since $E[Y|\boldsymbol{x}]$ and $\hat{Y}$ are constant given $\boldsymbol{x}$. We further have

$$\begin{aligned}
E[f(\boldsymbol{x})-\widehat{Y}(\boldsymbol{x}))^2|\boldsymbol{x}] &= E(f(\boldsymbol{x})-E[\widehat{Y}(\boldsymbol{x})]+E[\widehat{Y}(\boldsymbol{x})]-\widehat{Y}(\boldsymbol{x}))^2|\boldsymbol{x}] \\
&= E(f(\boldsymbol{x})-E[\widehat{Y}(\boldsymbol{x})])^2|\boldsymbol{x}]+E[(E[\widehat{Y}(\boldsymbol{x})]-\widehat{Y}(\boldsymbol{x}))^2|\boldsymbol{x}]+ \\
&\quad 2E[(f(\boldsymbol{x})-E[\widehat{Y}(\boldsymbol{x})])(E[\widehat{Y}(\boldsymbol{x})]-\widehat{Y}(\boldsymbol{x}))|\boldsymbol{x}] \\
&= (f(\boldsymbol{x})-E[\hat{f}(\boldsymbol{x})|\boldsymbol{x}])^2+\mathrm{Var}[\hat{f}(\boldsymbol{x})|\boldsymbol{x}]
\end{aligned}$$

The first term is the bias (squared) while the second term is the variance, showing the bias-variance trade off.

(b) For a very restrictive estimator we will expect a high bias but a low variance

For a very flexible estimator we will expect a low bias but a high variance.

(c) We could use the same data to both fit and estimate the performance, but then we would underestimate the error.

We could leave out a test set for evaluation. We would then get an unbiased estimate of the error, but we have less data to fit.

We can do cross-validation in which case we obtain both large training sets and a large test set. The main disadvantage here is that it can be time-consuming. However, in many model some can utilize efficient algorithms for calculating this.

Exercise 10

(a) We have

$$\begin{aligned}\Pr(S|V) =& \frac{\Pr(S)\Pr(V|S)}{\Pr(S)\Pr(V|S)+\Pr(R)\Pr(V|R)} = \frac{(1-r)q}{(1-r)q+rp},\\ \Pr(R|V) =& 1-\Pr(S|V) = \frac{rp}{(1-r)q+rp}\end{aligned}$$

and

$$\begin{aligned}\Pr(S|V) >& 0.5\\ \Updownarrow&\\ \frac{\Pr(S|V)}{1-\Pr(S|V)} >& 1\\ \Updownarrow&\\ \frac{(1-r)q}{rp} >& 1\\ \Updownarrow&\\ \frac{q}{p} >& \frac{r}{1-r}\end{aligned}$$

that is if we observe V and $\frac{q}{p} > \frac{r}{1-r}$ we classify it as spam.

Similarly

$$\begin{aligned}\Pr(S|V^c) =& \frac{\Pr(S)\Pr(V^c|S)}{\Pr(S)\Pr(V^c|S)+\Pr(R)\Pr(V^c|R)} = \frac{(1-r)(1-q)}{(1-r)(1-q)+r(1-p)}\\ \Pr(R|V^c) =& 1-\Pr(S|V^c) = \frac{r(1-p)}{(1-r)(1-q)+r(1-p)}\end{aligned}$$

and

$$\begin{aligned}\Pr(S|V^c) >& 0.5\\ \Updownarrow&\\ \frac{\Pr(S|V^c)}{\Pr(R|V^c)} >& 1\\ \Updownarrow&\\ \frac{(1-r)(1-q)}{r(1-p)} >& 1\\ \Updownarrow&\\ \frac{1-q}{1-p} >& \frac{r}{1-r}\end{aligned}$$

that is if we observe V^c and $\frac{1-q}{1-p} > \frac{r}{1-r}$ we classify to spam.

(b) We can introduce c_R to be a measure on the loss for classifying wrongly a real mail to spam and similarly c_S a measure on the loss of classifying a spam mail to be a real mail. Typically we will have $c_R > c_S$.

Expected loss is then

$$E[L] = E[E[L|x]]$$

where

$$\begin{aligned} E[L|x] =& c_R \Pr(Y = R, \widehat{Y} = S|x) + c_S \Pr(Y = S, \widehat{Y} = R|x) \\ =& c_R[1 - \Pr(Y = S|x)]I(\widehat{Y} = S) + c_S 1 \Pr(Y = S|x)I(\widehat{Y} = R) \end{aligned}$$

This indicate that in order to miminize the loss we should put $\widehat{Y} = S$ if $c_R[1-\Pr(Y = s|x)] < c_S \Pr(Y = S|x)$ which is equivalent to $\Pr(Y = s|x) > c_R/(c_R + c_S)$.

(c) It depends on independence of the presence of these words.

$$\begin{aligned} \Pr(S|\boldsymbol{V}) =& \frac{\Pr(S)\Pr(\boldsymbol{V}|S)}{\Pr(S)\Pr(\boldsymbol{V}|S) + \Pr(R)\Pr(\boldsymbol{V}|R)} \\ =& \frac{(1-r)\prod_{m=1}^{M} q_m^{V_m}(1-q_m)^{1-V_m}}{(1-r)\prod_{m=1}^{M} q_m^{V_m}(1-q_m)^{1-V_m} + r\prod_{m=1}^{M} p_m^{V_m}(1-p_m)^{1-V_m}} \end{aligned}$$

$$\begin{aligned} \frac{\Pr(S|V)}{\Pr(R|V)} >& 1 \\ \Updownarrow& \\ \frac{(1-r)\prod_{m=1}^{M} q_m^{V_m}(1-q_m)^{1-V_m}}{r\prod_{m=1}^{M} p_m^{V_m}(1-p_m)^{1-V_m}} >& 1 \\ \Updownarrow& \\ \frac{\prod_{m=1}^{M} q_m^{V_m}(1-q_m)^{1-V_m}}{\prod_{m=1}^{M} p_m^{V_m}(1-p_m)^{1-V_m}} >& \frac{r}{1-r} \end{aligned}$$

(d) The advantage of looking at pair of words is that it can be easier to obtain the whole meaning of the mail. The disadvantage is that there will be more parameters to estimate and that in short mails such pairs will occur very rare.