ETH Zurich
D-MATH
Dr. C. Buser
Prof. Dr. M. Wüthrich **January 24, 2019**

Examination

Data Analytics for Non-Life Insurance Pricing

Please fill in the following table

Last name	
First name	
Programme of study	**MATH** □ **SAV** □ **Other** □
Matriculation number	

Leave blank

Question	Maximum	Points	Check
1	10		
2	10		
3	10		
4	10		
Total	**40**		

Instructions

Duration of exam: 120 min.

Closed book examination: no notes, no books, no calculator, no smartphones, etc., allowed.

Important:

- ◇ Please put your student card (or an identification card for SAV students) on the table.
- ◇ Only pen and paper are allowed on the table. Please do **not** write with a **pencil** or a **red** or **green** pen. Moreover, please do not use **whiteout**.
- ◇ Start by reading all questions and answer the ones which you think are easier first, before proceeding to the ones you expect to be more difficult. Do not spend too much time on one question but try to solve as many questions as possible.
- ◇ Take a new sheet for each question and write your name on every sheet.
- ◇ All results have to be **explained/argued** by indicating intermediate steps in the respective calculations. You can use known formulas from the lecture without derivation.
- ◇ Simplify your results as far as possible.
- ◇ Some of the subquestions can be solved independently of each other.

⋆⋆⋆ Good luck! ⋆⋆⋆

Question 1 (**10 points**)

Assume we have $n = 10$ observations given by $\mathcal{D} = \{(Y_1, \boldsymbol{x}_1, v_1), \ldots, (Y_n, \boldsymbol{x}_n, v_n)\}$. Assume that Y_i are independent and Poisson distributed for $i = 1, \ldots, n$ with

$$Y_i \sim \text{Poisson}\left(\lambda(\boldsymbol{x}_i) v_i\right),$$

for given volumes $v_i = 1$ and with (unknown) regression function $\boldsymbol{x} \mapsto \lambda(\boldsymbol{x}) > 0$.

(a) Assume that the features $\boldsymbol{x}_i$ are (continuous) real-valued for all $i = 1, \ldots, n$. We have collected the following data $\mathcal{D}$:

i	1	2	3	4	5	6	7	8	9	10
Y_i	10	10	10	20	20	20	10	10	10	10
$\boldsymbol{x}_i$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
v_i	1	1	1	1	1	1	1	1	1	1

(i) How many (non-trivial) splits will the standardized binary regression tree perform if we use the Poisson deviance statistics as loss function? How big is the final Poisson deviance statistics loss? Give arguments for your answers.

(ii) Which is the first split that the standardized binary regression tree will make under the Poisson deviance statistics loss function? Give the right guess!

(b) Assume that the features $\boldsymbol{x}_i$ are (unordered) categorical for all $i = 1, \ldots, n$. We have collected the following explicit data $\mathcal{D}$:

i	1	2	3	4	5	6	7	8	9	10
Y_i	10	10	10	20	20	20	10	10	10	10
$\boldsymbol{x}_i$	BE	VS	TG	GR	AI	SG	TI	UR	SZ	LU
v_i	1	1	1	1	1	1	1	1	1	1

(i) How many (non-trivial) splits will the standardized binary regression tree perform if we use the Poisson deviance statistics as loss function? How big is the final Poisson deviance statistics loss? Give arguments for your answers.

(ii) Which is the first split that the standardized binary regression tree will make under the Poisson deviance statistics loss function?

(c) Prove that every standardized binary split considered in items (a) and (b) strictly decreases the Poisson deviance statistics loss.

(d) Comparing the results of items (a) and (b) we obtain the following in-sample losses on $\mathcal{D}$ and out-of-sample losses on test data $\mathcal{T}$.

	in-sample loss	out-of-sample loss
homogeneous model $\lambda(\cdot) \equiv$ constant	14.9630	19.9202
continuous features and 1 split	10.1939	9.4271
categorical features and 1 split	0.0000	15.4518

Discuss the two error measures (in-sample loss and out-of-sample loss) and make a model choice (with justification).

Solution 1

(a) (i) We see that the three observations $(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3) = (0.1, 0.2, 0.3)$ have the same response ($Y_i = 10$). Likewise, the three observations $(\boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}_6) = (0.4, 0.5, 0.6)$ have the same response ($Y_i = 20$). Finally, also the four observations $(\boldsymbol{x}_7, \boldsymbol{x}_8, \boldsymbol{x}_9, \boldsymbol{x}_{10}) = (0.7, 0.8, 0.9, 1)$ have the same response ($Y_i = 10$). This implies that on each of the three blocks of observations mentioned above we can get a perfect fit to the data. We conclude that the standardized binary regression tree will perform two (non-trivial) splits (one between 0.3 and 0.4, and the other between 0.6 and 0.7) and that the final Poisson deviance statistics loss is equal to 0.

(ii) If we first split the data between 0.6 and 0.7, then the Poisson deviance statistics loss is already equal to 0 for the four data points $(\boldsymbol{x}_7, \boldsymbol{x}_8, \boldsymbol{x}_9, \boldsymbol{x}_{10})$. Moreover, the regression tree estimator for the observations $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_6)$ is equal to $(3 \cdot 10 + 3 \cdot 20)/6 = 15$, i.e. it is exactly the average of the two responses 10 and 20 observed for the observations $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_6)$. Therefore, we guess that the first split will be done between 0.6 and 0.7. We remark that this is just a heuristic guess. One would need to calculate the Poisson deviance statistics loss in order to mathematically determine the first split.

(b) (i) The seven observations $(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_7, \boldsymbol{x}_8, \boldsymbol{x}_9, \boldsymbol{x}_{10}) = (\mathrm{BE}, \mathrm{VS}, \mathrm{TG}, \mathrm{TI}, \mathrm{UR}, \mathrm{SZ}, \mathrm{LU})$ have the same response ($Y_i = 10$). Likewise, the three observations $(\boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}_6) = (\mathrm{GR}, \mathrm{AI}, \mathrm{SG})$ have the same response ($Y_i = 20$). This implies that on each of these two blocks of observations we can get a perfect fit to the data. We conclude that the standardized binary regression tree will perform only one (non-trivial) split (one leaf will contain the observations $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_7, \boldsymbol{x}_8, \boldsymbol{x}_9, \boldsymbol{x}_{10}\}$ and the other leaf the observations $\{\boldsymbol{x}_4, \boldsymbol{x}_5, \boldsymbol{x}_6\}$) and that the final Poisson deviance statistics loss is equal to 0.

(ii) As we will only have one (non-trivial) split in this case, it is automatically clear what the first split will be, and we do not need to make a guess.

(c) Suppose that the considered split divides the leaf $\mathcal{X}_t$ into the two parts $\mathcal{X}_{t0}$ and $\mathcal{X}_{t1}$. Let $\mathcal{D}_t^*$ be the deviance statistics of the leaf $\mathcal{X}_t$ (where the split will happen). Moreover, let $\mathcal{D}_{t0}^*$ and $\mathcal{D}_{t1}^*$ be the deviance statistics after the split on the two resulting leaves $\mathcal{X}_{t0}$ and $\mathcal{X}_{t1}$. Then, we have

$$\begin{aligned}
\mathcal{D}_t^* &= \min_\lambda \sum_{i \in \mathcal{X}_t} 2Y_i \left[\frac{\lambda v_i}{Y_i} - 1 - \log\left(\frac{\lambda v_i}{Y_i} \right) \right] \\
&\geq \min_\lambda \sum_{i \in \mathcal{X}_{t0}} 2Y_i \left[\frac{\lambda v_i}{Y_i} - 1 - \log\left(\frac{\lambda v_i}{Y_i} \right) \right] + \min_\lambda \sum_{i \in \mathcal{X}_{t1}} 2Y_i \left[\frac{\lambda v_i}{Y_i} - 1 - \log\left(\frac{\lambda v_i}{Y_i} \right) \right] \qquad (1) \\
&= \mathcal{D}_{t0}^* + \mathcal{D}_{t1}^*.
\end{aligned}$$

As the MLE in the Poisson case is unique and the estimates of λ change in all the (non-trivial) splits considered in items (a) and (b), in the inequality in equation (1) we have in fact a "$>$" instead of a "$\geq$".

(d) The in-sample loss is the loss obtained on the training sample that is used to fit the model. Minimizing this error can lead to overfitting, especially if our considered model is too flexible. Therefore, the quality of the model should be evaluated on a test data set (out-of-sample loss) that has not been used for model estimation.

In our case we prefer the continuous feature model (with 1 split), because the corresponding model has the smallest out-of-sample loss (9.4271).

Question 2 (**10 points**)

Assume we have n observations given by

$$\mathcal{D} = \{Y_1, \dots, Y_n\}.$$

Assume that Y_i are independent and exponentially distributed for $i = 1, \dots, n$ with density

$$f_Y(y) \;=\; \theta \exp\{-\theta y\} \;\mathbb{1}_{\{y \geq 0\}},$$

for a given (but unknown) parameter $\theta > 0$.

(a) Calculate the maximum likelihood estimator $\widehat{\theta}^{\text{MLE}}$ for θ, given data $\mathcal{D}$, under the above model assumptions.

(b) Define a Bayesian model for the estimation of the unknown parameter θ using a non-degenerate prior distribution.
Hint: The gamma distribution has density supported on $\mathbb{R}_+$ and given by

$$\pi(\theta) = \frac{c^\gamma}{\Gamma(\gamma)} \theta^{\gamma-1} \exp\{-c\theta\} \;\mathbb{1}_{\{\theta>0\}}, \tag{2}$$

with given parameters $\gamma, c > 0$. The corresponding mean and variance are given by γ/c and γ/c^2, respectively.

(c) Calculate the posterior estimator $\widehat{\theta}^{\text{post}}$ for θ, given data $\mathcal{D}$, under the Bayesian model assumptions using π given in (2) as prior density.

(d) Give a credibility theory interpretation of the posterior estimator $\widehat{\theta}^{\text{post}}$ derived in item (c). Which is the parameter driving prior uncertainty if we assume that the prior mean $\theta_0 = \gamma/c$ is a given constant? Give an argument for your answer.

(e) Derive the (conditional) mean square error of prediction of $\widehat{\theta}^{\text{post}}$ derived in item (c). What happens with this error if $n \to \infty$? Give an argument for your answer.

Solution 2

(a) Assuming that $Y_i \geq 0$ for all $i = 1, \dots, n$ (which holds $\mathbb{P}$-a.s.), the likelihood function $L_{\mathcal{D}}(\theta)$ of the data $\mathcal{D}$ is given by

$$L_{\mathcal{D}}(\theta) \;=\; \prod_{i=1}^{n} \theta \exp\{-\theta Y_i\}.$$

Thus, for the log-likelihood we get

$$l_{\mathcal{D}}(\theta) \;\overset{\text{def}}{=}\; \log(L_{\mathcal{D}}(\theta)) \;=\; \sum_{i=1}^{n} \log(\theta) - \theta Y_i.$$

In order to determine the maximum likelihood estimator $\widehat{\theta}^{\text{MLE}}$ for θ, we take the derivative of $l_{\mathcal{D}}(\theta)$ with respect to θ and set it equal to 0. We have

$$\frac{\partial l_{\mathcal{D}}(\theta)}{\partial \theta} \;=\; \sum_{i=1}^{n} \frac{1}{\theta} - Y_i,$$

which is equal to 0 if and only if

$$n\frac{1}{\theta} = \sum_{i=1}^{n} Y_i \quad \Longleftrightarrow \quad \theta = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} Y_i}.$$

For the second derivative of $l_{\mathcal{D}}(\theta)$ we get

$$\frac{\partial^2 l_{\mathcal{D}}(\theta)}{\partial \theta^2} \;=\; -\frac{n}{\theta^2} < 0.$$

We conclude that the log-likelihood function is concave in θ, and that the maximum likelihood estimator $\widehat{\theta}^{\text{MLE}}$ is given by

$$\widehat{\theta}^{\text{MLE}} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} Y_i}.$$

(b) In a Bayesian model we assume that the parameter θ is a random variable whose density π is supported on $(0, \infty)$. Additionally, we assume that

$$Y_i \mid \theta \sim \text{exponential}(\theta)$$

for all $i = 1, \dots, n$ and that, conditionally on θ, the random variables $Y_1, \dots, Y_n$ are independent.

Using the definition of the conditional density (Bayes' theorem) and our assumptions, the joint distribution of the data $\boldsymbol{Y} = (Y_1, \dots, Y_n)'$ and the parameter θ is given by the density

$$f(\boldsymbol{Y}, \theta) = f(\boldsymbol{Y}|\theta)\pi(\theta) = \left(\prod_{i=1}^{n} f(Y_i|\theta)\right)\pi(\theta),$$

where $f(\boldsymbol{Y}|\theta)$ denotes the conditional density of $\boldsymbol{Y}$ given θ and, analogously, $f(Y_i|\theta)$ denotes the conditional density of Y_i given θ. The posterior distribution of θ is then the distribution of θ given the data $\boldsymbol{Y}$, and is given by the density

$$f(\theta|\boldsymbol{Y}) = \frac{f(\boldsymbol{Y}, \theta)}{f(\boldsymbol{Y})} \propto f(\boldsymbol{Y}|\theta)\pi(\theta).$$

(c) In order to identify the posterior distribution, we select for the prior distribution of θ the gamma distribution given in the hint in item (b). In that case, again assuming that $Y_i \geq 0$ for all $i = 1, \dots, n$, the posterior distribution of the parameter θ given the data $\boldsymbol{Y}$ is given by

$$\begin{aligned} f(\theta|\boldsymbol{Y}) &\propto \left(\prod_{i=1}^{n} \theta \exp\{-\theta Y_i\} \right) \frac{c^\gamma}{\Gamma(\gamma)} \theta^{\gamma-1} \exp\{-c\theta\} \\ &\propto \theta^{\gamma+n-1} \exp\left\{ - \left(c + \sum_{i=1}^{n} Y_i \right) \theta \right\}, \end{aligned}$$

which is the unnormalized density of the gamma distribution with parameters

$$\widehat{\gamma}^{\text{post}} = \gamma + n \quad \text{and} \quad \widehat{c}^{\text{post}} = c + \sum_{i=1}^{n} Y_i.$$

Since we have $\widehat{\theta}^{\text{post}} = \mathbb{E}[\theta|\boldsymbol{Y}]$, using again the hint from item (b), we obtain

$$\widehat{\theta}^{\text{post}} = \frac{\widehat{\gamma}^{\text{post}}}{\widehat{c}^{\text{post}}} = \frac{\gamma + n}{c + \sum_{i=1}^{n} Y_i}.$$

(d) We can write

$$\widehat{\theta}^{\text{post}} = \frac{\gamma + n}{c + \sum_{i=1}^{n} Y_i} = \frac{\gamma}{c} \frac{c}{c + \sum_{i=1}^{n} Y_i} + \frac{n}{\sum_{i=1}^{n} Y_i} \frac{\sum_{i=1}^{n} Y_i}{c + \sum_{i=1}^{n} Y_i} = (1 - w)\, \theta_0 + w\, \widehat{\theta}^{\text{MLE}},$$

where θ_0 is the mean of the prior distribution π, $\widehat{\theta}^{\text{MLE}}$ is the MLE from item (a) and w is the credibility weight given by

$$w = \frac{\sum_{i=1}^{n} Y_i}{c + \sum_{i=1}^{n} Y_i} \in (0, 1).$$

If we assume that the prior mean $\theta_0 = \gamma/c$ is a given constant, then, using the hint in item (b), we get

$$\text{Var}(\theta) = \frac{\gamma}{c^2} = \frac{\theta_0}{c}.$$

In particular, we see that the parameter c drives the prior uncertainty (for fixed prior mean $\theta_0 = \gamma/c$).

(e) We have that

$$\text{MSE}\left(\widehat{\theta}^{\text{post}} \middle| \boldsymbol{Y} \right) = \mathbb{E}\left[\left(\widehat{\theta}^{\text{post}} - \theta \right)^2 \middle| \boldsymbol{Y} \right] = \mathbb{E}\left[\left(\mathbb{E}\left[\theta|\boldsymbol{Y}\right] - \theta \right)^2 \middle| \boldsymbol{Y} \right] = \text{Var}\left(\theta|\boldsymbol{Y}\right).$$

Using again the hint from item (b), we obtain

$$\begin{aligned} \text{MSE}\left(\widehat{\theta}^{\text{post}} \middle| \boldsymbol{Y} \right) &= \frac{\widehat{\gamma}^{\text{post}}}{(\widehat{c}^{\text{post}})^2} = \frac{\gamma + n}{(c + \sum_{i=1}^{n} Y_i)^2} = \frac{\gamma + n}{c + \sum_{i=1}^{n} Y_i} \frac{1}{c + \sum_{i=1}^{n} Y_i} \\ &= \frac{\frac{\gamma}{n} + 1}{\frac{c}{n} + \frac{1}{n} \sum_{i=1}^{n} Y_i} \frac{1}{c + \sum_{i=1}^{n} Y_i}. \end{aligned}$$

Due to the law of large numbers, we have

$$\frac{1}{n} \sum_{i=1}^{n} Y_i \longrightarrow \mu, \text{ for some } \mu \in (0, \infty), \qquad \text{and} \qquad \sum_{i=1}^{n} Y_i \to \infty, \qquad \text{as } n \to \infty.$$

We can conclude that $\text{MSE}\left(\widehat{\theta}^{\text{post}} \middle| \boldsymbol{Y} \right) \to 0$, as $n \to \infty$.

Question 3 (10 points)

Assume we have n claims given by

$$\mathcal{D} = \{(Y_1, \boldsymbol{x}_1), \ldots, (Y_n, \boldsymbol{x}_n)\}.$$

Assume that $\boldsymbol{x}_i \in \mathcal{X} = \mathbb{R}$, and that Y_i are independent and log-normally distributed for $i = 1, \ldots, n$ with density supported in $\mathbb{R}_+$ and given by

$$Y_i \sim f(y|\boldsymbol{x}_i) = \frac{1}{\sqrt{2\pi}} \frac{1}{y} \exp\left\{ -\frac{1}{2} \left(\log y - \mu(\boldsymbol{x}_i)\right)^2 \right\}, \qquad \text{for } y \geq 0,$$

for a given (but unknown) regression function $\mu : \mathcal{X} \to \mathbb{R}$.

(a) Calculate the deviance statistics for this problem using a general regression function $\mu : \mathcal{X} \to \mathbb{R}$.

(b) Set up a single hidden layer neural network with 10 hidden neurons for this regression problem, using the hyperbolic tangent activation function given by $\phi(x) = (e^x - e^{-x})/(e^x + e^{-x})$ for $x \in \mathbb{R}$. Calculate the number of parameters of this model.

(c) Calculate one step of the gradient descent optimization algorithm explicitly for the deviance statistics loss function derived in item (a) and the single hidden layer neural network defined in item (b). Why is the hyperbolic tangent an attractive activation function in the application of the gradient descent algorithm?

(d) Choose the neural network defined in item (b) but replace the hyperbolic tangent activation function by the step function activation $\phi(x) = \mathbb{1}_{\{x \geq 0\}}$ for $x \in \mathbb{R}$.

(i) Can the number of parameters of this regression function be reduced compared to the one in item (b) without affecting the regression function itself? If yes, how many parameters are sufficient? Justify your answer.

(ii) How many different output values $\mu(\boldsymbol{x})$ can this regression model at most produce? Justify your answer.

(e) Assume that the feature space $\mathcal{X}$ is categorical having 11 different labels, i.e. the feature space is given by $\mathcal{X} = \{a1, a2, a3, a4, a5, a6, a7, a8, a9, a10, a11\}$. We use dummy coding for these categorical feature components, and then we set up a single hidden layer neural network having one hidden neuron and hyperbolic tangent activation function.

(i) Calculate the number of parameters that this regression model receives.

(ii) How many different output values $\mu(\boldsymbol{x})$ can this regression model at most produce? Justify your answer.

Solution 3

(a) Assuming that $Y_i \geq 0$ for all $i = 1, \ldots, n$ (which holds $\mathbb{P}$-a.s.), the likelihood function $L_{\mathcal{D}}(\mu(\cdot))$ of the data $\mathcal{D}$ is given by

$$L_{\mathcal{D}}(\mu(\cdot)) \;=\; \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \frac{1}{Y_i} \exp\left\{ -\frac{1}{2}(\log(Y_i) - \mu(\boldsymbol{x}_i))^2 \right\}.$$

Thus, for the log-likelihood we get

$$l_{\mathcal{D}}(\mu(\cdot)) \;=\; \log(L_{\mathcal{D}}(\mu(\cdot))) \;=\; \sum_{i=1}^{n} -\log\left(\sqrt{2\pi}\right) - \log(Y_i) - \frac{1}{2}(\log(Y_i) - \mu(\boldsymbol{x}_i))^2.$$

In the saturated model we have one parameter μ_i per observation Y_i. That is, we have to maximize

$$g(\mu_i) \overset{\text{def}}{=} -\log\left(\sqrt{2\pi}\right) - \log(Y_i) - \frac{1}{2}(\log(Y_i) - \mu_i)^2$$

with respect to μ_i, for all $i = 1, \ldots, n$. If we take the derivative with respect to μ_i, we get

$$\frac{\partial g(\mu_i)}{\partial \mu_i} = \log(Y_i) - \mu_i,$$

for all $i = 1, \ldots, n$. This is equal to 0 if and only if

$$\widehat{\mu}_i = \widehat{\mu}_i(Y_i) = \log(Y_i), \tag{3}$$

for all $i = 1, \ldots, n$. For the second derivative of $g(\mu_i)$ with respect to μ_i we get

$$\frac{\partial^2 g(\mu_i)}{\partial \mu_i^2} = -1 < 0,$$

for all $i = 1, \ldots, n$. That is, in the saturated model we have the parameter $\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\mu}}(\boldsymbol{Y}) = (\widehat{\mu}_1(Y_1), \ldots, \widehat{\mu}_n(Y_n))$ with $\widehat{\mu}_i(Y_i)$ given as in (3), for all $i = 1, \ldots, n$. For the log-likelihood of the saturated model we then have

$$\begin{aligned} l_{\mathcal{D}}(\widehat{\boldsymbol{\mu}}) \;&=\; \sum_{i=1}^{n} -\log\left(\sqrt{2\pi}\right) - \log(Y_i) - \frac{1}{2}(\log(Y_i) - \log(Y_i))^2 \\ &=\; \sum_{i=1}^{n} -\log\left(\sqrt{2\pi}\right) - \log(Y_i). \end{aligned}$$

Finally, the (scaled) deviance statistics is given by

$$\begin{aligned} D^*(\boldsymbol{Y}, \mu(\cdot)) &= 2(l_{\mathcal{D}}(\widehat{\boldsymbol{\mu}}) - l_{\mathcal{D}}(\mu(\cdot))) \\ &= 2\sum_{i=1}^{n} -\log\left(\sqrt{2\pi}\right) - \log(Y_i) + \log\left(\sqrt{2\pi}\right) + \log(Y_i) + \frac{1}{2}(\log(Y_i) - \mu(\boldsymbol{x}_i))^2 \\ &= \sum_{i=1}^{n} (\log(Y_i) - \mu(\boldsymbol{x}_i))^2. \end{aligned}$$

(b) We choose a single hidden layer neural network with 10 hidden neurons. As our feature space is $\mathcal{X} = \mathbb{R}$, we have only one neuron in the input layer. Using the hyperbolic tangent activation function ϕ given on the exam sheet, we have activations, for all $j = 1, \ldots, 10$,

$$z_j(x) = \phi(w_{j,0} + w_{j,1}x),$$

with unknown parameters $w_{j,0}, w_{j,1} \in \mathbb{R}$, for the 10 neurons in the hidden layer. Since the codomain of $\mu(\cdot)$ is the real line, we define a linear regression approach as follows

$$\mu(x) = \beta_0 + \sum_{j=1}^{10} \beta_j z_j(x),$$

with unknown parameters $\beta_0, \beta_1, \ldots, \beta_{10} \in \mathbb{R}$. Overall, we have

$$(1+1)10 + (10+1) = 31$$

parameters in the model.

(c) We note that for the derivative of the hyperbolic tangent activation function ϕ we have

$$\frac{\partial \phi(x)}{\partial x} = \frac{\partial}{\partial x} \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{(e^x + e^{-x})^2 - (e^x - e^{-x})^2}{(e^x + e^{-x})^2} = 1 - \phi^2(x). \tag{4}$$

We write

$$\boldsymbol{\theta} = (w_{1,0}, w_{1,1}, \ldots, w_{10,0}, w_{10,1}, \beta_0, \beta_1, \ldots, \beta_{10}) \in \mathbb{R}^{31}$$

for the vector of the unknown model parameters. Thus, the regression function $\mu_{\boldsymbol{\theta}}(\cdot)$ depends on $\boldsymbol{\theta}$. In the gradient descent optimization algorithm the goal is to decrease a given loss function by iteratively updating the model parameters. In our case we would like to decrease the deviance statistics

$$D^*(\boldsymbol{Y}, \mu_{\boldsymbol{\theta}}(\cdot)) = \sum_{i=1}^{n} (\log(Y_i) - \mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i))^2.$$

To this end, for a given $\boldsymbol{\theta}$, we move in the direction of the maximal local decrease of the deviance statistics, i.e. in the direction of the negative gradient $\nabla_{\boldsymbol{\theta}} D^*(\boldsymbol{Y}, \mu_{\boldsymbol{\theta}}(\cdot))$ of the deviance statistics. We calculate

$$\nabla_{\boldsymbol{\theta}} D^*(\boldsymbol{Y}, \mu_{\boldsymbol{\theta}}(\cdot)) = \frac{\partial D^*(\boldsymbol{Y}, \mu_{\boldsymbol{\theta}}(\cdot))}{\partial \boldsymbol{\theta}} = -2 \sum_{i=1}^{n} (\log(Y_i) - \mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i)) \frac{\partial \mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i)}{\partial \boldsymbol{\theta}},$$

where we have

$$\begin{aligned}
\frac{\partial \mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i)}{\partial w_{j,0}} &= \beta_j (1 - z_j^2(\boldsymbol{x}_i)), \\
\frac{\partial \mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i)}{\partial w_{j,1}} &= \beta_j \boldsymbol{x}_i (1 - z_j^2(\boldsymbol{x}_i)), \\
\frac{\partial \mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i)}{\partial \beta_0} &= 1, \\
\frac{\partial \mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i)}{\partial \beta_j} &= z_j(\boldsymbol{x}_i),
\end{aligned}$$

for all $i = 1, \ldots, n$ and $j = 1, \ldots, 10$. In one single step of the gradient descent optimization algorithm we have the update

$$\boldsymbol{\theta} \longrightarrow \boldsymbol{\theta} - \rho \nabla_{\boldsymbol{\theta}} D^*(\boldsymbol{Y}, \mu_{\boldsymbol{\theta}}(\cdot)),$$

where $\rho > 0$ is the so-called learning rate. The hyperbolic tangent activation function is an attractive activation function in the application of the gradient descent algorithm because of its property (4). This property allows to efficiently calculate gradients, as it only requires simple subtraction and multiplication and as there is no need for re-evaluating some other function. Moreover, with the hyperbolic tangent activation function the activations in the neurons in the hidden layers lie in $[-1, 1]$. This can considerably speed up training, as there is no risk of the activations to explode. This holds especially true for deep neural networks.

(d) (i) With the step function activation $\phi(x) = \mathbb{1}_{\{x \geq 0\}}$, for $x \in \mathbb{R}$, the regression function $\mu(\cdot)$ looks as follows

$$\mu(x) = \beta_0 + \sum_{j=1}^{10} \beta_j \phi(w_{j,0} + w_{j,1}x) = \beta_0 + \sum_{j=1}^{10} \beta_j \mathbb{1}_{\{w_{j,0}+w_{j,1}x \geq 0\}}.$$

Note that wlog we can assume that $w_{j,1} > 0$, for all $j = 1, \ldots, 10$: If $w_{k,1} < 0$ for some $k \in \{1, \ldots, 10\}$, then we can rewrite $\mu(\cdot)$ as follows

$$\begin{aligned}\mu(x) &= \beta_0 + \sum_{\substack{j=1\\ j\neq k}}^{10} \beta_j \mathbb{1}_{\{w_{j,0}+w_{j,1}x \geq 0\}} + \beta_k \mathbb{1}_{\{w_{k,0}+w_{k,1}x \geq 0\}} \\ &= (\beta_0 + \beta_k) + \sum_{\substack{j=1\\ j\neq k}}^{10} \beta_j \mathbb{1}_{\{w_{j,0}+w_{j,1}x \geq 0\}} - \beta_k \mathbb{1}_{\{-w_{k,0}-w_{k,1}x \geq 0\}}.\end{aligned}$$

If $w_{k,1} = 0$ for some $k \in \{1, \ldots, 10\}$, then the regression function does not depend on the value of this particular neuron, i.e. the considered neuron could be removed from the model. Thus, we assume that $w_{j,1} > 0$, for all $j = 1, \ldots, 10$. Then, the regression function $\mu(\cdot)$ can be rewritten as

$$\mu(x) = \beta_0 + \sum_{j=1}^{10} \beta_j \mathbb{1}_{\{w_{j,0}+w_{j,1}x \geq 0\}} = \beta_0 + \sum_{j=1}^{10} \beta_j \mathbb{1}_{\left\{x \geq -\frac{w_{j,0}}{w_{j,1}}\right\}} = \beta_0 + \sum_{j=1}^{10} \beta_j \mathbb{1}_{\{x \geq \widetilde{w}_j\}},$$

where $\widetilde{w}_j = -\frac{w_{j,0}}{w_{j,1}}$, for all $j = 1, \ldots, 10$. In particular, all tuples $(w_{j,0}, w_{j,1})$ can be replaced by one parameter $\widetilde{w}_j$. This corrresponds to a reduction of one parameter for all $j = 1, \ldots, 10$, leading to a new total of $31 - 10 = 21$ parameters.

(ii) As the parameters $\widetilde{w}_j$ can be ordered as $\widetilde{w}_{(1)} \leq \widetilde{w}_{(2)} \leq \cdots \leq \widetilde{w}_{(10)}$, the regression function $\mu(\cdot)$ is actually a step function which is constant in $[-\infty, \widetilde{w}_{(1)}]$, in $[\widetilde{w}_{(j)}, \widetilde{w}_{(j+1)}]$, for all $j = 1, \ldots, 9$, and in $[\widetilde{w}_{(10)}, \infty]$. Hence, the regression function $\mu(\cdot)$ can produce at most 11 different values (namely in the case, where the parameters $\widetilde{w}_1, \ldots, \widetilde{w}_{10}$ are all different from each other).

(e) (i) If we use dummy coding for a categorical variable having 11 different labels, we need $11 - 1 = 10$ variables. For a given feature $a \in \mathcal{X}$ we define

$$u_l(a) = \begin{cases} 1, & \text{if } a = a_l, \\ 0, & \text{else,} \end{cases}$$

for all $l = 1, \ldots, 10$. The label a_{11} then corresponds to the vector

$$(u_1(a_{11}), \ldots, u_{10}(a_{11})) = (0, \ldots, 0)$$

of all zeroes, i.e. label a_{11} is chosen as reference label. The regression function $\mu(\cdot)$ is given by

$$\mu(a) = \beta_0 + \beta_1 \phi\left(w_{1,0} + \sum_{l=1}^{10} w_{1,l}\, u_l(a)\right),$$

where $\phi(\cdot)$ is the hyperbolic tangent activation function. We see that the regression function $\mu(\cdot)$ depends on 13 parameters $(\beta_0, \beta_1, w_{1,0}, w_{1,1}, \ldots, w_{1,10})$.

(ii) For a given feature $a \in \mathcal{X}$, either all the $u_l(a)$ are 0 or exactly one of them is equal to 1. This leads to at most 11 different values inside the hyperbolic tangent activation function, which, in turn, gives rise to at most 11 different values of the regression function $\mu(\cdot)$.

Question 4 (10 points)

Assume we have n independent and identically distributed claims count observations given by the data

$$\mathcal{D} = \{(N_1, \boldsymbol{x}_1), \ldots, (N_n, \boldsymbol{x}_n)\}.$$

(a) What is the advantage of a log-linear regression function structure in terms of model interpretation of the coefficients considered? Give a short proof of your statement for the following regression function

$$\log \lambda(x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d. \tag{5}$$

(b) You have the feature component `age` as a continuous variable in the regression function together with other continuous variables and categorical factors:

$$\log \lambda(x) = \beta_0 + \beta_1 \texttt{age} + \beta_2 x_2 + \ldots + \beta_d x_d.$$

You expected the variable `age` to be highly significant, but it is not. What could be the problem and how can you solve it? Give 2 possible solutions.

(c) One of your explanatory variables (feature component x_1) shows the following problem. The variable is significant in a model without the other explanatory variables:

$$\log \lambda(x) = \beta_0 + \beta_1 x_1.$$

But in a model with all explanatory variables given by (5) the significance of feature component x_1 vanishes while many other variables are significant. What is a possible reason? Would you use the variable x_1 for your final tariff? Give an argument for your decision.

(d) (i) Assume you have a categorical explanatory variable with many levels (lots of them with only few observations). How would you use this variable in your generalized linear model? Give 2 possibilities with their advantages and disadvantages.

(ii) Assume you have several continuous explanatory variables and you do not know the best functional form to include them in the generalized linear model. What can you do? Give 2 possibilities with their advantages and disadvantages.

Solution 4

(a) A log-linear regression function leads to a multiplicative structure. That is, for the feature vector $x = (x_1, \dots, x_d)$ we have

$$\lambda(x) = \exp\{\beta_0\} \exp\{\beta_1 x_1\} \cdots \exp\{\beta_d x_d\}.$$

Let us now assume that the (continuous) feature value x_1 gets changed by the amount Δx. The feature vector is then given by $\tilde{x} = (x_1 + \Delta x, x_2, \dots, x_d)$ and the resulting regression function gets affected in a multiplicative way:

$$\lambda(\tilde{x}) = \exp\{\beta_0\} \exp\{\beta_1 (x_1 + \Delta x)\} \cdots \exp\{\beta_d x_d\} = \lambda(x) \exp\{\beta_1 \Delta x\}.$$

(b) It is still possible that the variable `age` is highly significant. The proposed model suggests that the logarithmic response is linearly dependent on the variable `age`. However, if in reality we observe that the logarithmic response is e.g. a quadratic function of the variable `age`, then a simple log-linear model cannot capture this kind of dependence structure, resulting in a variable `age` which is not significative. To detect the dependence structure between the response and the variable `age`, one should consider a marginal plot. In case of the aforementioned quadratic dependence one can replace the variable `age` by the variable $\texttt{age}^2$.

On a completely different note, the effect of the variable `age` on the response could be masked by a correlated feature, leading to collinearity problems. See also item (c) below.

(c) A possible reason for this phenomenon is collinearity, i.e. there could be another feature component (e.g. feature component x_2), which is correlated with feature component x_1. In this case it can happen that the feature component x_2 is significative and x_1 is not significative anymore, as its effect on the response variable is already explained by x_2. There are arguments for keeping the variable x_1 in the model as well as there are arguments for removing the variable x_1 from the model. An argument for keeping the variable x_1 would be that it is still possible that x_1 is a model-relevant variable and removing it would result in a bias. An argument for removing the variable x_1 would be that a model should be as sparse as possible in order to not overfit to the data and keep the variance of the coefficients estimates as low as possible. As the variable x_1 shows itself not to be significative in the joint model, it is a good candidate to be removed from the model. In the end it really depends on the variable under consideration whether to keep it in the model or remove it.

(d) (i) In order to use such an explanatory variable with many levels in a generalized linear model, one could try to combine levels. This could be done according to the logic of the feature, e.g. zip codes could be summarized to the various regions, areas or districts of a city or a country. Where such an aggregation is not possible, the levels could also be combined according to their response rates, i.e. we combine those levels where we observe similar average responses. The advantage of such aggregation techniques is that we do not have as many levels as before, and that we have more observations for the remaining levels, which allows us to build a more robust model. On the other hand, the model gets more crude, loosing additional granular information that one could have gained with the original categorical variable if one had had enough data.

As another possibility, the variables could be directly replaced by their response rate, transforming the categorical variable into a continuous one. Here the advantage is that a continuous variable can directly be used as a feature and no additional data pre-processing methods such as dummy coding are necessary. The disadvantage here is that one gives up the true categorical nature of the feature and that one introduces continuity where there might be none. Moreover, if one has only few observation for a given level, then the corresponding response rate might not be very representative. Consequently, there is a risk of introducing a bias into the model. Note that this drawback holds also true for the aggregation of feature levels on the basis of the response rate explained above.

(ii) One could for example study the marginal plots of the response variable against each of the continuous explanatory variables. This allows us to get an idea of how the response is related to the features. However, one has also to be careful as the marginal plots do not tell us how the features interact amongst each other, leading to possibly different functional forms in the joint model.

As a second possibility one could move from the generalized linear models towards the generalized additive models. The advantage is that a generalized additive model setup allows to build a more flexible model structure. However, a more flexible model needs greater care, especially during the fitting procedure. Moreover, generalized additive models tend to be harder to interpret than generalized linear models.

By choosing a generalized additive model we stay in the world of models with an additive/multiplicative structure. Testing interactions could be done by regression trees, boosting or neural networks.