ETH Zurich
D-MATH
Dr. C. Buser
Prof. Dr. M. Wüthrich

August 21, 2018

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 observations given by

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

Assume that Y_i are independent and Bernoulli distributed for $i = 1, \ldots, n$ with

$$Y_i = \begin{cases} 1 & \text{with probability } p(\boldsymbol{x}_i), \\ 0 & \text{with probability } 1 - p(\boldsymbol{x}_i), \end{cases}$$

for a given (but unknown) regression function $p : \mathcal{X} \to (0, 1)$.

(a) Choose a homogeneous regression function, i.e. $p(\boldsymbol{x}) \equiv p \in (0, 1)$ for all $\boldsymbol{x} \in \mathcal{X}$. Give the resulting log-likelihood function and derive the maximum likelihood estimator $\widehat{p}$ for p.

(b) Calculate the resulting in-sample deviance statistics and give sufficient conditions for the observations $\mathcal{D}$ such that the resulting estimated distribution is non-degenerate.

(c) Assume that $\boldsymbol{x} \in \mathcal{X}$ is a one-dimensional continuous real-valued feature, i.e. $\mathcal{X} = \mathbb{R}$. Define a generalized linear model for the estimation of the regression function $p : \mathcal{X} \to (0, 1)$ using 4 (non-empty) categorical classes. Calculate the resulting maximum likelihood estimator. *Hint:* Use for data compression in the categorical classes the property that the sum of i.i.d. Bernoulli distributed random variables provides a random variable with a well-known distribution function.

(d) Assume that $\boldsymbol{x} \in \mathcal{X}$ is a one-dimensional continuous real-valued feature, i.e. $\mathcal{X} = \mathbb{R}$. Define a generalized linear model for the estimation of the regression function $p : \mathcal{X} \to (0, 1)$ directly using the continuous feature $\boldsymbol{x}$. Give the design matrix and calculate the resulting maximum likelihood estimator (as far as possible).

(e) Comparing the results of items (a), (c) and (d) we obtain the following in-sample losses and out-of-sample losses.

	in-sample loss	out-of-sample loss
(a) homogeneous model	0.2320	0.2360
(c) categorical feature	0.2050	0.2200
(d) continuous feature	0.2100	0.2180

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

Question 2 (10 points)

Assume we have n observations given by

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

Assume that Y_i are independent and Bernoulli distributed for $i = 1, \ldots, n$ with

$$Y_i = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1-p, \end{cases}$$

for a given (but unknown) parameter $p \in (0,1)$.

(a) Define a Bayesian Bernoulli model for the estimation of the unknown parameter $p \in (0,1)$ using a non-degenerate prior distribution.
Hint: The Beta distribution has density supported on $(0,1)$ given by

$$\pi(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \qquad \text{for } y \in (0,1),$$

and given parameters $\alpha, \beta > 0$. The corresponding mean and variance are given by $\alpha/(\alpha+\beta)$ and $\alpha\beta/((\alpha+\beta)^2(\alpha+\beta+1))$, respectively.

(b) Calculate the posterior estimator $\widehat{p}^{\text{post}}$ for p, given data $\mathcal{D}$, under the Bayesian model assumptions made in item (a) (using π as prior density).

(c) Give a credibility theory interpretation of the posterior estimator $\widehat{p}^{\text{post}}$ derived in the previous item. Can the posterior estimator lead to a degenerate probability model under the above model assumptions (give an argument for your answer)?

(d) Derive the (conditional) mean square error of prediction of $\widehat{p}^{\text{post}}$ derived under item (b). What happens with this error if $n \to \infty$?

(e) Explain why this Bayesian Bernoulli model can be useful in regression tree constructions.

Question 3 (10 points)

Assume we have n large 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 Pareto distributed for $i = 1, \ldots, n$ with density supported in $[M, \infty)$ and given by

$$Y_i \sim f(y|\boldsymbol{x}_i) = \frac{\alpha(\boldsymbol{x}_i)}{M} \left(\frac{y}{M}\right)^{-\alpha(\boldsymbol{x}_i)-1}, \qquad \text{for } y \geq M,$$

for a given (known) large claims threshold $M > 0$ and a given (but unknown) regression function $\alpha : \mathcal{X} \to \mathbb{R}_+$.

(a) Calculate the deviance statistics for this problem.

(b) Set up a single hidden layer neural network with more than two hidden neurons for this regression problem using the sigmoid activation function. How many parameters does the model have?

(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). Explain why the gradient descent method is of interest in neural network calibrations.

(d) Assume we have a large number of hidden neurons (say more than 100). Why are we in this situation in general not interested in finding the maximum likelihood estimator? What alternative solution do you propose?

(e) Assume we have feature space $\mathcal{X} = [-1, 1]^2$. Compare a single hidden layer neural network with 3 hidden neurons and step function activation to a gradient boosting machine, where for the latter we use single split regression trees for totally 3 boosting steps. Which of the two models has the smaller optimal in-sample loss (give an argument for your answer)? Which of the two models has the smaller out-of-sample loss (give an argument for your answer)?

Question 4 (10 points)

Assume we have n independently distributed claims count observations given by the data

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

An actuary wants to have your opinion based on the following output. Take your decisions on a test level of $\alpha = 5\%$.

(a) Define an appropriate generalized linear model for claims frequency modeling based on the given data $\mathcal{D}$. What conditions need to be fulfilled so that the model can be applied?

(b) The actuary gives you the following R output of his analysis. Answer the following questions based on his output:

(i) How many observations do we have?

(ii) How many explanatory variables are available and what structure do they have?

(iii) Based on the output below: which variables have a significant relationship with the observed claims frequency? Give statistical arguments for your statements.

```
> summary(regr2 <- glm(formula = N ~ f1, family = quasipoisson(link =
                  "log"), data = dat))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4567  -1.4093  -1.3660   0.0686   6.6602

Coefficients:
```

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.07	0.05	-1.5	0.14450	
f1B	0.13	0.07	2.0	0.04850	*
f1C	0.06	0.07	0.9	0.34630	

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for for quasipoisson family taken to be 2.105839)

    Null deviance: 5486.7  on 2999  degrees of freedom
Residual deviance: 5478.5  on 2997  degrees of freedom
AIC: NA

> drop1(regr2, test="Chisq")
Single term deletions

Model:
N ~ f1
```

	Df	Deviance	scaled dev.	Pr(>Chi)
<none>		5'478.5		
f1	2	5'486.7	3.9096	0.14160

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> exp(coef(regr2))
```

(Intercept)	f1B	f1C
0.9	1.15	1.05

(c) Assume that you have decided to keep variable f1 in the model. Give the resulting prediction for the claims frequencies of the different policies.

(d) You intend to improve your existing generalized linear model and want to keep the interpretability at the same time.

(i) What could you do to improve the prediction of your existing model?

(ii) How would you compare different models to check which model performs better?

(e) Consider the following output below and compare it to the output from item (b).

(i) What are the differences between the two models?

(ii) Would you revise one or more statements that you have taken in item (b) based on this new output? Give arguments for your statements.

(iii) Which model fits better to the data? Give arguments for your statements.

```
> summary(regr1 <- glm(formula = N ~ f1, family = poisson(link =
        "log"), data = dat))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4567  -1.4093  -1.3660   0.0686   6.6602

Coefficients:
```

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-0.07	0.03	-2.1	0.03415	*
f1b	0.13	0.04	2.9	0.00418	**
f2c	0.06	0.05	1.4	0.17164	

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 5486.7  on 2999  degrees of freedom
Residual deviance: 5478.5  on 2997  degrees of freedom
AIC: 9123.2

> drop1(regr1, test="Chisq")
Single term deletions

Model:
N ~ f1
```

	Df	Deviance	AIC	LRT	Pr(>Chi)	
<none>		5'478.5	9'123.2			
f1	2	5'486.7	9'127.5	8.233	0.01630	*

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> exp(coef(regr1))
```

(Intercept)	f1B	f1C
0.9	1.15	1.05