

Statistical Methods - III partial test

January 10, 2024 - 1.15 hour

To allow evaluation, you must submit the text with name and surname.

Name and Surname	Student ID	1
------------------	------------	---

1 Code 1 reports the summary of a fitted Gaussian GAM on y , with covariates x_0 , x_1 , x_2 , x_3 . Which of the following statements is true?

- ☐ It is not sensible to consider smooth effects for x_0 , x_1 and x_2
- ☐ The coefficient for x_3 is statistically significant at 5%
- ☐ The R-squared is 0.724
- ☐ The total estimated degrees of freedom is 27

2 Figure 1 shows the predicted surface of a model regressing the city distance on the curb weight and the engine size. Which of the following models could allow at obtaining such a surface?

- ☐ Ridge regression
- ☐ GAM
- ☐ Linear model
- ☐ Regression tree

3 Which of the following statements on regression splines represented by the truncated-power basis functions is true?

- ☐ In a cubic regression spline setting the number of knots and their position are automatically selected
- ☐ Regression splines do not suffer of high uncertainty at the boundary regions
- ☐ Parameter estimation is carried out using penalised least squares
- ☐ The number of parameters to be estimated for a quadratic regression spline on the single predictor x considering 2 internal knots is 5

4 Consider the random sample, $\mathbf{y} = (0, 0, 0, 0, 1)$ of size 5 from $\text{Be}(p)$. By considering a prior $p \sim \text{Beta}(\alpha, \beta)$, with $\alpha = 2$ and $\beta = 4$ which is the correct posterior distribution?

- ☐ $p|\mathbf{y} \sim \text{Beta}(\alpha^*, \beta^*)$, with $\alpha^* = 3$, $\beta^* = 8$
- ☐ $p|\mathbf{y} \sim \text{Beta}(\alpha^*, \beta^*)$, with $\alpha^* = 8$, $\beta^* = 3$
- ☐ $p|\mathbf{y} \sim \text{Beta}(\alpha^*, \beta^*)$, with $\alpha^* = 4$, $\beta^* = 9$
- ☐ $p|\mathbf{y} \sim \text{Beta}(\alpha^*, \beta^*)$, with $\alpha^* = 6$, $\beta^* = 5$

5 Which of the following statements on LASSO regression is false?

- ☐ The regression coefficient estimator is biased but has smaller variance than the least squares estimator
- ☐ The penalised likelihood framework allows to shrink the coefficient estimates towards zero
- ☐ The penalisation term is responsible for variable selection
- ☐ The regression coefficient estimator can be expressed in a closed form

6 Which of the following statements on the ensemble methods is true?

- ☐ Bagging and Boosting can only consider decision trees as weak learners
- ☐ Boosting (more specifically AdaBoost) enhances Bagging by replacing an equal-weight random sampling of the units with an adaptive weighted-sampling procedure mainly targeting poorly predicted/classified observations
- ☐ Procedures for statistical inference, such as confidence intervals and hypothesis testing, are easily available
- ☐ Bagging, Boosting and Random Forest are only suited for classification tasks

7 Given the following dataset $\mathbf{x} = (31, 14, 56, 24, 43, 28, 45, 25, 13, 42)$, which of the following vectors can not be a bootstrap replicate?

- ☐ $\mathbf{x}^* = (43, 43, 43, 43, 43, 43, 43, 43, 43, 43)$
☐ $\mathbf{x}^* = (31, 14, 56, 24, 30, 28, 45, 25, 13, 42)$
☐ $\mathbf{x}^* = (31, 14, 56, 24, 43, 28, 45, 25, 13, 42)$
☐ $\mathbf{x}^* = (42, 31, 14, 42, 25, 45, 14, 42, 31, 31)$

8 Code 1 reports the summary of a fitted Gaussian GAM on y , with covariates x_0, x_1, x_2, x_3 . By denoting with λ_j the parameter controlling the balance between smoothness and model fit in estimation associated with the effect of x_j , which of the following answers is a sensible estimate of the smoothing parameter vector $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \lambda_2)$?

- ☐ (0.022, 0.004, 1.98) ☐ (0.022, 1.98, 0.004) ☐ (1.98, 0.022, 0.004) ☐ (19800, 0.022, 0.004)

9 Which of the following statements on decision trees is false?

- ☐ Pruning a decision tree allows to reduce the variance at the cost of a little bias
☐ The decision tree algorithm considers a cost complexity term in the objective function in the growing and pruning phases
☐ Growing a tree and pruning it could allow for variable selection
☐ Full growing a decision tree results in high variance and low bias

10 A study would confirm the effectiveness of a new drug. In a previous test the drug is seen to be effective on the 90 percent of patients. Which prior should be favoured to include such an information in a Bayesian analysis workflow?

- ☐ $\mathcal{N}(0.9, 1)$ ☐ $\text{Unif}(0, 1)$ ☐ $\text{Beta}(5, 45)$ ☐ $\text{Beta}(45, 5)$

Code 1 - GAM output

```
Family: gaussian
Link function: identity

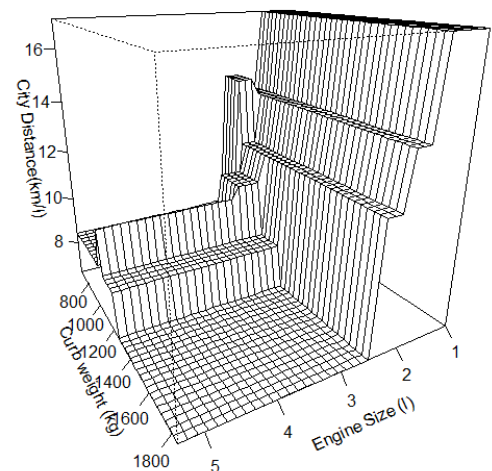
Formula:
y ~ s(x0) + s(x1) + s(x2) + x3

Parametric coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.0841     0.2028  39.871  <2e-16 ***
x3            -0.3006     0.3532  -0.851   0.395
---

Approximate significance of smooth terms:
              edf Ref.df      F p-value
s(x0)  2.771   3.448  7.978 2.1e-05 ***
s(x1)  6.916   7.996 37.515 < 2e-16 ***
s(x2)  8.474   8.919 75.205 < 2e-16 ***
---

R-sq.(adj) =  0.71   Deviance explained = 72.4%
GCV = 4.5263   Scale est. = 4.2982    n = 400
```

Figure 1



Statistical Methods - III partial test

January 10, 2024 - 1.15 hour

To allow evaluation, you must submit the text with name and surname.

Name and Surname	Student ID	2
------------------	------------	---

1 Which of the following statements on decision trees is false?

- ☐ The decision tree algorithm considers a cost complexity term in the objective function in the growing and pruning phases
- ☐ Growing a tree and pruning it could allow for variable selection
- ☐ Full growing a decision tree results in high variance and low bias
- ☐ Pruning a decision tree allows to reduce the variance at the cost of a little bias

2 Consider the random sample, $\mathbf{y} = (1, 1, 1, 1, 0)$ of size 5 from $\text{Be}(p)$. By considering a prior $p \sim \text{Beta}(\alpha, \beta)$, with $\alpha = 4$ and $\beta = 2$ which is the correct posterior distribution?

- ☐ $p|\mathbf{y} \sim \text{Beta}(\alpha^*, \beta^*)$, with $\alpha^* = 3$, $\beta^* = 8$
- ☐ $p|\mathbf{y} \sim \text{Beta}(\alpha^*, \beta^*)$, with $\alpha^* = 9$, $\beta^* = 4$
- ☐ $p|\mathbf{y} \sim \text{Beta}(\alpha^*, \beta^*)$, with $\alpha^* = 5$, $\beta^* = 6$
- ☐ $p|\mathbf{y} \sim \text{Beta}(\alpha^*, \beta^*)$, with $\alpha^* = 8$, $\beta^* = 3$

3 A study would confirm the effectiveness of a new drug. In a previous test the drug is seen to be effective on the 90 percent of patients. Which prior should be favoured to include such an information in a Bayesian analysis workflow?

- ☐ $\text{Unif}(0, 1)$
- ☐ $\text{Beta}(5, 45)$
- ☐ $\text{Beta}(45, 5)$
- ☐ $\mathcal{N}(0.9, 1)$

4 Code 1 reports the summary of a fitted Gaussian GAM on $\mathbf{y}$, with covariates $\mathbf{x}_0$, $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$. Which of the following statements is true?

- ☐ The coefficient for $\mathbf{x}_3$ is statistically significant at 5%
- ☐ The R-squared is 0.724
- ☐ The total estimated degrees of freedom is 27
- ☐ It is not sensible to consider smooth effects for $\mathbf{x}_0$, $\mathbf{x}_1$ and $\mathbf{x}_2$

5 Given the following dataset $\mathbf{x} = (31, 14, 56, 24, 43, 28, 45, 25, 13, 42)$, which of the following vectors can not be a bootstrap replicate?

- ☐ $\mathbf{x}^* = (31, 14, 56, 24, 30, 28, 45, 25, 13, 42)$
- ☐ $\mathbf{x}^* = (31, 14, 56, 24, 43, 28, 45, 25, 13, 42)$
- ☐ $\mathbf{x}^* = (45, 31, 24, 24, 28, 42, 31, 25, 28, 25)$
- ☐ $\mathbf{x}^* = (43, 43, 43, 43, 43, 43, 43, 43, 43, 43)$

6 Code 1 reports the summary of a fitted Gaussian GAM on $\mathbf{y}$, with covariates $\mathbf{x}_0$, $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$. By denoting with λ_j the parameter controlling the balance between smoothness and model fit in estimation associated with the effect of $\mathbf{x}_j$, which of the following answers is a sensible estimate of the smoothing parameter vector $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \lambda_2)$?

- ☐ (0.022, 1.98, 0.004)
- ☐ (1.98, 0.022, 0.004)
- ☐ (19800, 0.022, 0.004)
- ☐ (0.022, 0.004, 1.98)

7 Which of the following statements on ridge regression is false?

- ☐ The penalised likelihood framework allows to shrink the coefficient estimates towards zero
- ☐ The penalisation term is responsible for variable selection
- ☐ The regression coefficient estimator can be expressed in a closed form
- ☐ The regression coefficient estimator is biased but has smaller variance than the least squares estimator

8 Which of the following statements on regression splines represented by the truncated-power basis functions is true?

- ☐ Regression splines do not suffer of high uncertainty at the boundary regions
- ☐ Parameter estimation is carried out using penalised least squares
- ☐ The number of parameters to be estimated for a linear regression spline on the single predictor x considering 3 internal knots is 5
- ☐ In a cubic regression spline setting the number of knots and their position are automatically selected

9 Which of the following statements on the ensemble methods is true?

- ☐ Boosting (more specifically AdaBoost) enhances Bagging by replacing an equal-weight random sampling of the units with an adaptive weighted-sampling procedure mainly targeting poorly predicted/classified observations
- ☐ Procedures for statistical inference, such as confidence intervals and hypothesis testing, are easily available
- ☐ Bagging, Boosting and Random Forest are only suited for classification tasks
- ☐ Bagging and Boosting can only consider decision trees as weak learners

10 Figure 1 shows the predicted surface of a model regressing the city distance on the curb weight and the engine size. Which of the following models could not allow at obtaining such a surface?

- ☐ GAM
- ☐ Linear model
- ☐ Regression tree
- ☐ Ridge regression

Code 1 - GAM output

```
Family: gaussian
Link function: identity

Formula:
y ~ s(x0) + s(x1) + s(x2) + x3

Parametric coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.0841     0.2028  39.871  <2e-16 ***
x3            -0.3006     0.3532  -0.851    0.395
---

Approximate significance of smooth terms:
              edf Ref.df    F p-value
s(x0)  2.771   3.448   7.978 2.1e-05 ***
s(x1)  6.916   7.996  37.515 < 2e-16 ***
s(x2)  8.474   8.919  75.205 < 2e-16 ***
---

R-sq.(adj) =  0.71   Deviance explained = 72.4%
GCV = 4.5263   Scale est. = 4.2982    n = 400
```

Figure 1

