Exam

University of Twente, EEMCS

Statistical Learning 2024

November 7, 2024, 8.45-11.45

Grade = $1 + 9 \times \text{points}/30$.

It is allowed to use a basic, non-graphical calculator. Please write with a pen (not a pencil). No other tools are allowed, in particular no smart tools such as phones or smart watches.

Part 1: Theory

1. (3 points) The Gamma distribution with parameters $\alpha, \beta > 0$ is given by the probability density function

$$f(x|\alpha,\beta) = \frac{1}{\Gamma(\alpha)}\beta^\alpha x^{\alpha-1}e^{-\beta x}\mathbf{1}(x>0),$$

with $\Gamma(\alpha) := \int_0^\infty x^{\alpha-1}e^{-x}dx$. Suppose we observe N i.i.d. copies from the Gamma distribution with parameters $\alpha = 1$ and $\beta > 0$. Let the prior distribution on β be a Gamma distribution with parameters $\nu, \gamma > 0$. Determine the posterior distribution of β.

2. (3 points) Prove that if a minimax rule is unique, it is admissibe.

3. (3 points) Suppose we tested $d = 5$ hypotheses $H_1, \ldots, H_5$ which yielded the p-values $0.019, 0.105, 0.015, 0.045,$ and 0.304 respectively. Use the method of Benjamini-Hochberg to control the FDR at level 5%, and indicate which hypotheses are to be rejected (if your p-value is exactly equal to the critical value, you may choose to reject it).

4. Recall that the p.d.f. of an exponential distributed random variable with parameter $\theta > 0$ (denoted in the following by $\text{Exp}(\theta)$) is given by $\theta^{-1}\exp(-x/\theta)\mathbf{1}(x > 0)$. It is well known that the life span distribution of many items (mobile phones, cars, light bulbs, ...) follows an exponential distribution. Suppose we observe for n mobile phones whether they still work after a fixed time $s > 0$ and we are interested in recovering the parameter θ. Thus, in this case, the full (unobserved) dataset consists of n i.i.d. exponential random variables, that is, $X_1, \ldots, X_n \sim \text{Exp}(\theta)$ modeling the life spans of the n mobile phones. The observed data are $S_i = \mathbf{1}(X_i \geq s)$, $i = 1, \ldots, n$. This means $S_i = 1$ if the i-th mobile phone still works after time s and $S_i = 0$ otherwise.

   a) (1 point) Derive the log-likelihood of the full data model, where we observe $X_1, \ldots, X_n$.

   b) (1 point) Show that the p.d.f. of $X_i|(X_i \geq s)$ is $x \mapsto \theta^{-1}\exp((\overset{s}{t} - x)/\theta)\mathbf{1}(x \geq \overset{s}{t})$.

   c) (1 point) Show that the p.d.f. of $X_i|(X_i < s)$ is

$$x \mapsto \frac{e^{-x/\theta}}{\theta(1-e^{-s/\theta})}\mathbf{1}\big(0 \leq x < s\big).$$

   You can use that for any real number a, $\int_a^\infty ue^{-u}du = (a+1)e^{-a}$.

   d) (1 point) Prove that for any $\theta' > 0$,

$$E_{\theta'}\big[X_i|(X_i \geq s)\big] = s + \theta'.$$

   e) (1 point) Prove that for any $\theta' > 0$,

$$E_{\theta'}\big[X_i|(X_i < s)\big] = \theta' - \frac{se^{-s/\theta'}}{1-e^{-s/\theta'}}.$$

f) (2 points) Show that the E-step in the EM-algorithm is given by

$$-n\log\theta - \frac{\theta^{(t)}n}{\theta} - \frac{s}{\theta}\sum_{i=1}^{n} S_i + \frac{se^{-s/\theta^{(t)}}}{\theta(1-e^{-s/\theta^{(t)}})}\Big(n - \sum_{i=1}^{n} S_i\Big).$$

g) (2 points) Derive the M-step (that is, the update for $\theta^{(t+1)}$) of the EM-algorithm.

Part 2: concepts

The following questions are multiple choice. Only write down the number of the question and the letter(s) of the answer(s). Note that to each question multiple answers can (but don't need to) be correct!

Each question is worth 2 points. For each correct answer you get (sub)points. For each wrong answer you get minus points (but never ending up below 0). Example: Answer A and B are correct. A student answers A, B, C and D. This yields 2 points for A and B but 2 minus points for C and D, resulting in 0 points for that question.

---

5. Which of the following statements about Bayesian inference is/are true?
   a) Subjective Bayesians think that prior probabilities should be rationally constrained in a way that a single rule determines a unique prior for every situation.
   b) One of the upshots of Bayesian inference is that all inference is done via the posterior distribution.
   c) The more data we obtain, the less important the choice of prior distribution becomes.
   d) It is possible to use an improper prior distribution, yet to obtain a tractable posterior distribution.

6. What is/are the primary purpose(s) of the bootstrap method?
   a) To reduce overfitting by using kernel functions.
   b) To estimate the accuracy and variance of a model by generating multiple datasets from the original dataset.
   c) To improve model performance by increasing the size of the original dataset through data augmentation.
   d) To select the most relevant features by eliminating redundant ones.

7. What is/are (a) key characteristic(s) of Jeffreys' prior?
   a) It assigns equal probability to all possible parameter values
   b) It is an objective prior, invariant under reparameterisation
   c) It maximises the posterior probability given the data
   d) It is based on the subjective beliefs of the researcher

8. A computer learns to play the game of Go by watching 100 games of human players and learning about the outcome. Then, the computer plays a game. This is:
   a) Supervised learning
   b) Semi-supervised learning

c) Unsupervised learning

d) batch learning

e) online learning

f) active learning

g) passive learning

9. Which of the following statements about cross-validation and the jackknife resampling method are true?

   a) Both cross-validation and the jackknife involve splitting the data into multiple subsets, but cross-validation uses multiple random splits, while the jackknife systematically leaves out one observation at a time.

   [✓ b)] Cross-validation aims to assess model performance by using multiple train-test splits, while the jackknife is mainly used for estimating the bias and variance of a statistical estimator.

   c) Cross-validation is used primarily to estimate the accuracy of a predictive model, while the jackknife is used to generate new data points for training.

   d) The jackknife method requires multiple train-test splits, similar to cross-validation, but focuses on minimizing bias rather than variance.

   e) Cross-validation and the jackknife are both methods that use all available data for training without creating any test subsets.

10. Which of the following best describes the purpose of Markov Chain Monte Carlo (MCMC) methods?

   (a) MCMC methods are used to directly solve complex integrals analytically by finding exact solutions for probability distributions.

   (b) MCMC methods are primarily used for optimizing deterministic functions to find global maxima or minima.

   (c) MCMC methods aim to reduce overfitting in machine learning models by performing cross-validation on the training data.

   (d) MCMC methods generate samples from a probability distribution when direct sampling is difficult, allowing estimation/approximation of expectations and integrals.