EXAM IN STATISTICAL MACHINE LEARNING
STATISTISK MASKININLÄRNING

DATE: March 13, 2024

RESPONSIBLE TEACHER: Jens Sjölund

NUMBER OF PROBLEMS: 5

ALLOWED AIDS: Calculator, mathematical handbooks

PRELIMINARY GRADES:

grade 3	23 points
grade 4	33 points
grade 5	43 points

Some general instructions and information:

- Your solutions should be given in English.
- Only write on one page of the paper.
- Write your exam code and a page number on all pages.
- Do not use a red pen.
- Use separate sheets of paper for the different problems (i.e. the numbered problems, 1–5).

With the exception of Problem 1, all your answers must be clearly motivated! A correct answer without a proper motivation will score zero points!

Good luck!

1. This problem is composed of 10 true-or-false statements. You only have to classify these as either `true` or `false`. For this problem *(only!)* no motivation is required. Each correct answer scores 1 point and each incorrect answer scores -1 point (capped at 0 for the whole problem). Answers left blank score 0 points.

   i. False. Boosting can be used with any weak classifier.

   ii. False.

   iii. False. If your model is *overfitting*, training on more data can be expected to improve performance.

   iv. True.

   v. False. Without proper care, machine learning methods are likely to propagate biases in the data.

   vi. True.

   vii. True.

   viii. False.

   ix. True.

   x. False. In bagging, you create multiple datasets by resampling *data points* with replacement.

   (10p)

2. (a) From

$$\frac{e^{\theta^T \mathbf{x}}}{1 + e^{\theta^T \mathbf{x}}} = r$$

it follows that

$$\rho_{\text{db}} = \frac{\log \frac{r}{1-r} - \theta_0}{\theta_1} = 0.86.$$

(b) The confusion matrix is given by

	$\hat{y} = 0$	$\hat{y} = 1$
y = 0	7	1
y = 1	1	3

The diagonal entries give the numbers of correctly classified samples of the respective classes. The bottom left entry gives the number of false negatives whereas the top right entry gives the number of false positives. The misclassification rate is $\frac{2}{12} = 0.17$.

(c) Looking at the dataset reveals that $\rho'_{\text{db}} \in [0.5, 0.6)$ results in zero false negatives. The corresponding threshold is determined as

$$r' = \frac{e^{\theta_0 + \theta_1 \rho'_{\text{db}}}}{1 + e^{\theta_0 + \theta_1 \rho'_{\text{db}}}} \in [0.080, 0.146).$$

The false positive rate changes to $\frac{3}{12} = 0.25$. This means that the company has to discard more functioning components.

(d) The binary logistic regression model gives the class probabilities

$$p(y = 1|\mathbf{x}) = \frac{e^{\alpha^T \mathbf{x}}}{1 + e^{\alpha^T \mathbf{x}}}$$

and

$$p(y = 0|\mathbf{x}) = 1 - p(y = 1|\mathbf{x}) = \frac{1}{1 + e^{\alpha^T \mathbf{x}}}.$$

Letting $M = 2$ in the multiclass logistic regression model we obtain the following expression for the class probabilities:

$$\begin{bmatrix} p(y = 1|\mathbf{x}) \\ p(y = 0|\mathbf{x}) \end{bmatrix} = \frac{1}{e^{\boldsymbol{\beta_1}^T \mathbf{x}} + e^{\boldsymbol{\beta_2}^T \mathbf{x}}} \begin{bmatrix} e^{\boldsymbol{\beta_1}^T \mathbf{x}} \\ e^{\boldsymbol{\beta_2}^T \mathbf{x}} \end{bmatrix}$$

Following the hint, we can show that adding a constant vector $\mathbf{c}$ to the parameter vectors does not change the class probabilities:

$$\frac{e^{(\beta_\mathbf{i}+\mathbf{c})^T\mathbf{x}}}{\sum_{k=1}^{K} e^{(\beta_\mathbf{k}+\mathbf{c})^T\mathbf{x}}} = \frac{e^{\mathbf{c}^T\mathbf{x}} e^{\beta_\mathbf{i}{}^T\mathbf{x}}}{\sum_{k=1}^{K} e^{\mathbf{c}^T\mathbf{x}} e^{\beta_\mathbf{k}{}^T\mathbf{x}}} = \frac{e^{\beta_\mathbf{i}{}^T\mathbf{x}}}{\sum_{k=1}^{K} e^{\beta_\mathbf{k}{}^T\mathbf{x}}}.$$

If we let $\mathbf{c} = -\beta_\mathbf{2}$ in the multiclass logistic regression model, we obtain:

$$\begin{bmatrix} p(y=1|\mathbf{x}) \\ p(y=0|\mathbf{x}) \end{bmatrix} = \frac{1}{e^{\beta_\mathbf{1}{}^T\mathbf{x}} + e^{\beta_\mathbf{2}{}^T\mathbf{x}}} \begin{bmatrix} e^{\beta_\mathbf{1}{}^T\mathbf{x}} \\ e^{\beta_\mathbf{2}{}^T\mathbf{x}} \end{bmatrix} =$$

$$= \frac{1}{e^{(\beta_{10}-\beta_{20})+(\beta_{11}-\beta_{21})\rho} + e^0} \begin{bmatrix} e^{(\beta_{10}-\beta_{20})+(\beta_{11}-\beta_{21})\rho} \\ e^0 \end{bmatrix} =$$

$$= \frac{1}{1 + e^{(\beta_{10}-\beta_{20})+(\beta_{11}-\beta_{21})\rho}} \begin{bmatrix} e^{(\beta_{10}-\beta_{20})+(\beta_{11}-\beta_{21})\rho} \\ 1 \end{bmatrix}.$$

Finally, if we let $\alpha_0 = \beta_{10} - \beta_{20}$ and $\alpha_1 = \beta_{11} - \beta_{21}$, we see that the predicted class probabilities in the multiclass logistic regression model are identical to the class probabilities obtained in the binary logistic regression model.

3. (a) One valid solution is given in Figure 1, but there may be others.

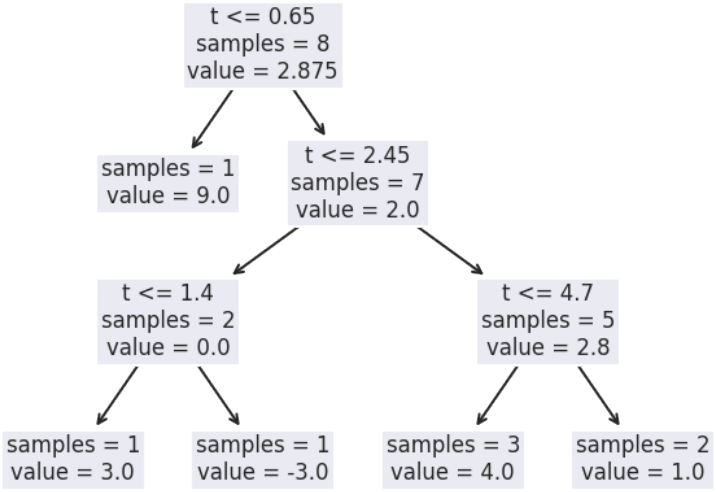


Figure 1: Regression tree of depth tree that solves question 3a.

From the figure in the question, we see that the trained tree has five constant regions and therefore must have (at least) five leaves. Since we make binary splits, a tree of depth d can have at most 2^d leaves. We can thus conclude that the minimal depth is three.

(b) We compute the mean-squared error on the *validation* data for each choice of k.

k = 1

$$f_{k=1}(0.9) = 9,$$
$$f_{k=1}(4.1) = 5$$
$$E_{\text{val},k=1} = \frac{(3-9)^2 + (4-5)^2}{2} = \frac{37}{2} = 18.5.$$

k = 2

$$f_{k=2}(0.9) = \frac{9-3}{2} = 3,$$
$$f_{k=2}(4.1) = \frac{5+1}{2} = 3,$$
$$E_{\text{val},k=2} = \frac{(3-3)^2 + (4-3)^2}{2} = 0.5.$$

k = 3

$$f_{k=3}(0.9) = \frac{9-3+5}{3} = 11/3,$$
$$f_{k=3}(4.1) = \frac{-3+5+1}{3} = 1,$$
$$E_{\text{val},k=3} = \frac{(3-\frac{11}{3})^2 + (4-1)^2}{2} = \frac{85}{18} \approx 4.7.$$

Clearly, the smallest validation error is attained for $k = 2$.
We then estimate the expected new data error using the *test* data.

$$f_{k=2}(3.0) = \frac{-3+5}{2} = 1,$$
$$f_{k=2}(5.3) = \frac{5+1}{2} = 3,$$
$$E_{\text{test},k=2} = \frac{(3-1)^2 + (1-3)^2}{2} = 4.$$

In other words, our estimate is $E_{\text{new}} \approx 4$.

(c) The measured voltage v is related to model $u(t)$ according to

$$v(t) = u(t) + b + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

Shortcut The quick route to solving the problem is to realize that the maximum likelihood solution is equivalent to a standard least-squares problem in a single variable. In particular, one can immediately realize that, since the mean is the minimizer of the sum of squares, the optimal choice is to set b equal to the mean of the residuals $v(t_i) - u(t_i)$. For completeness, however, we will now show how to find the solution using the standard maximum likelihood procedure.

Standard ML Let $u_i = u(t_i)$. From the equation above, it follows that the likelihood is

$$\begin{aligned}
\log \mathcal{L}(b) &= \log p(\mathbf{v} \,|\, b) \\
&= \log \mathcal{N}(\mathbf{v} \,|\, \mathbf{u} + b, \sigma^2) \\
&= \sum_{i=1}^{8} \log \mathcal{N}(v_i) \,|\, u_i + b, \sigma^2) \\
&\propto \frac{1}{2\sigma^2} \sum_{i=1}^{8} (v_i - u_i - b)^2
\end{aligned}$$

Taking the derivative with respect to b and setting it to zero gives

$$0 = \frac{d}{db} \log \mathcal{L}(b) = -\frac{1}{\sigma^2} \sum_{i=1}^{8} (v_i - u_i - b)$$

$$b = \frac{1}{8} \sum_{i=1}^{8} (v_i - u_i) \approx 2.12.$$

4. (a) Computing the parameters we obtain:

$$\hat{\pi}_1 = \frac{n_1}{N} = \frac{2}{5} \qquad \hat{\pi}_{-1} = \frac{n_{-1}}{N} = \frac{3}{5}$$
$$\hat{\mu}_1 = \frac{1}{n_1} \sum_{x_i : y_i = 1} x_i = [45, 0.7]^T$$
$$\hat{\mu}_{-1} = \frac{1}{n_{-1}} \sum_{x_i : y_i = -1} x_i = [20, 1.3]^T$$
$$\hat{\Sigma} = \begin{bmatrix} 50 & 0.6 \\ 0.6 & 0.056 \end{bmatrix}$$

   (b) In LDA, the joint probability is modeled as two factors: $p(y)$ and $p(x|y) = \frac{p(x,y)}{p(y)}$. To get $p(y, x)$, we first sample from $p(y)$ which is a categorical distribution over the class labels and then choose a sample from the modeled Gaussian $p(x|y)$.

   (c) Generative models model the joint probability distribution $p(x, y)$ of labels and data, while discriminative models are only modeling the conditional probability distribution $p(y|x)$.

   (d) The log-posterior of the model is given by

$$\begin{aligned} \log p(y = m|x) &\propto \log p(x|y = m) + \log p(y = m) \\ &= \log \mathcal{N}(x|\hat{\mu}_m, \hat{\Sigma}) + \log \hat{\pi}_m \\ &= \log \frac{1}{(2\pi)^{p/2}\sqrt{\det\hat{\Sigma}}} - \frac{1}{2}(x - \hat{\mu}_m)^\intercal \hat{\Sigma}^{-1}(x - \hat{\mu}_m) + \log \hat{\pi}_m \\ &\propto -\frac{1}{2}(x - \hat{\mu}_m)^\intercal \hat{\Sigma}^{-1}(x - \hat{\mu}_m) + \log \hat{\pi}_m. \end{aligned}$$

   Then we consider maximizing the log-posterior to obtain $\hat{y}_\star$:

$$\begin{aligned} \hat{y}_\star &= \arg\max_m \{p(y = m|x)\} \\ &= \arg\max_m \{\log p(y = m|x)\} \\ &= \arg\max_m \{-\frac{1}{2}(x - \hat{\mu}_m)^\intercal \hat{\Sigma}^{-1}(x - \hat{\mu}_m) + \log \hat{\pi}_m\}. \end{aligned}$$

5. (a) i. Regression, because the output is continuous (although restricted to be non-negative). (1p)

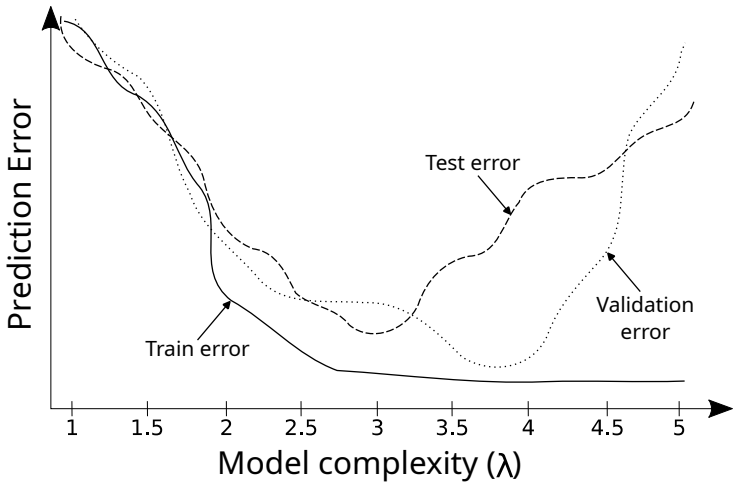
ii.

$$s^{(1)} = w^{(1)}x^{(0)} + b^{(1)} = \begin{pmatrix} 1.5 \\ 7.0 \\ -8.8 \end{pmatrix}, \quad x^{(1)} = \sigma(s^{(1)}) = \begin{pmatrix} 0.82 \\ 1.0 \\ 0.0 \end{pmatrix},$$

$$s^{(2)} = w^{(2)}x^{(1)} + b^{(2)} = \begin{pmatrix} -0.82 \end{pmatrix}, \quad x^{(2)} = \mathrm{ReLU}(s^{(2)}) = \begin{pmatrix} 0.0 \end{pmatrix}.$$

(3p)

iii. Learning rate, batch size, number of epochs. (2p)



(b) i. Within model selection, we choose the model that yields the lowest validation error. Hence, $\lambda \approx 3.8$ is best. (1p)

ii. We see underfitting for $\lambda < 2.5$ and overfitting for $\lambda > 4$. (2p)

iii. Use cross-validation. (1p)