EXAM IN STATISTICAL MACHINE LEARNING
STATISTISK MASKININLÄRNING

DATE: January 9, 2024

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. i) False.
   ii) False. [$L = -\ln p$]
   iii) False. [High variance.]
   iv) False. [Translational invariance.]
   v) True.
   vi) True.
   vii) True.
   viii) True.
   ix) True.
   x) False.

2. a) Using the squared error loss, we know from the fact sheet that

$$\widehat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \left(\begin{bmatrix}1 & 199.1\\ 1 & 228.5\end{bmatrix}^T \begin{bmatrix}1 & 199.1\\ 1 & 228.5\end{bmatrix}\right)^{-1} \begin{bmatrix}1 & 199.1\\ 1 & 228.5\end{bmatrix}^T \begin{bmatrix}6.60\\ 9.14\end{bmatrix} \approx \begin{bmatrix}-10.6\\ 0.09\end{bmatrix}$$

b) Predicted incubation time

$$f(\mathbf{x}; \widehat{\boldsymbol{\theta}}) = \mathbf{x}^T\widehat{\boldsymbol{\theta}} \approx -10.6 \cdot 1 + 0.09 \cdot 50 = -6.28$$

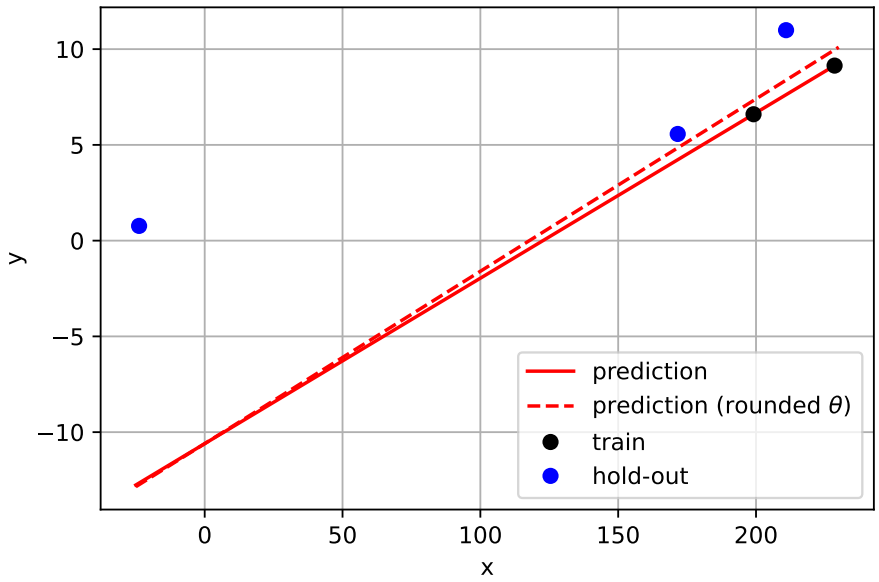
We actually obtain -6.28 using the computed $\widehat{\boldsymbol{\theta}}$ and -6.1 using $\widehat{\boldsymbol{\theta}}$ rounded to two digits after the decimal.

Note that the prediction isn't even in the range of positive incubation times!

c) Using the hold out set, we have the following estimate

$$E_{\text{hold-out}} = \frac{1}{3}\sum_{i=3}^{5}(y_i - f(\mathbf{x}; \widehat{\boldsymbol{\theta}}))^2 \approx 83.2$$

We actually obtain 64.49 using the computed $\widehat{\boldsymbol{\theta}}$ and 63.3 using $\widehat{\boldsymbol{\theta}}$ rounded to two digits after the decimal.



3. (a) We write out

$$J(\boldsymbol{\theta}, v) = \sum_{i=1}^{n} \frac{1}{2v}(\ln y_i - \mathbf{x}_i^T\boldsymbol{\theta})^2 + \frac{1}{2}\ln v + K_i = \frac{1}{2v}\|\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\theta}\|^2 + \frac{n}{2}\ln v + K$$

$$\begin{aligned}
J(\boldsymbol{\theta}, v) &= \ln \prod_{i=1}^{n} p(y_i|\mathbf{x}_i; \boldsymbol{\theta}, v) = \sum_{i=1}^{n} \ln p(y_i|\mathbf{x}_i; \boldsymbol{\theta}, v) \\
&= \sum_{i=1}^{n} \ln \left[\frac{1}{y_i\sqrt{2\pi v}} \exp\left(-\frac{(\ln y_i - \mathbf{x}_i^T\boldsymbol{\theta})^2}{2v} \right) \right] \\
&= \sum_{i=1}^{n} \ln 1 - \ln\left(y_i\sqrt{2\pi v}\right) - \frac{(\ln y_i - \mathbf{x}_i^T\boldsymbol{\theta})^2}{2v} \\
&= \sum_{i=1}^{n} \underbrace{-\ln y_i - \frac{1}{2}\ln \pi}_{=K_i} - \frac{1}{2}\ln v - \frac{1}{2v}(\ln y_i - \mathbf{x}_i^T\boldsymbol{\theta})^2 \\
&= -\frac{1}{2v}\sum_{i=1}^{n} \left[(\ln y_i - \mathbf{x}_i^T\boldsymbol{\theta})^2 \right] - \frac{n}{2}\ln v - \underbrace{\sum_{i=1}^{n} K_i}_{=K}
\end{aligned}$$

Instead of maximizing $J(\boldsymbol{\theta}, v)$, we multiply $J(\boldsymbol{\theta}, v)$ by (-1) and minimize it. Furthermore, note that the constant K can be dropped since it does neither depend on $\boldsymbol{\theta}$ nor v. Thus,

$$J(\boldsymbol{\theta}, v) = \frac{1}{2v}\|\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\theta}\|^2 + \frac{n}{2}\ln v$$

Thus minimization with respect to $\boldsymbol{\theta}$ is a equivalent to a least squares problem where

$$\widehat{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\widetilde{\mathbf{y}}, \quad \text{where } \widetilde{\mathbf{y}} \text{ is defined as in Task 2.}$$

Inserting this solution back we have

$$J(\widehat{\boldsymbol{\theta}}, v) = \frac{1}{2v}\|\widetilde{\mathbf{y}} - \mathbf{X}\widehat{\boldsymbol{\theta}}\|^2 + \frac{n}{2}\ln v + K$$

Now set the first derivative w.r.t. v to zero:

$$\partial_v J(\widehat{\boldsymbol{\theta}}, v) = -\frac{1}{2v^{-2}}\|\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\theta}\|^2 + \frac{n}{2v} = 0$$

$$\partial_v J(\widehat{\boldsymbol{\theta}}, v) = -\frac{1}{2v^2}\|\widetilde{\mathbf{y}} - \mathbf{X}\widehat{\boldsymbol{\theta}}\|^2 + \frac{n}{2v} \stackrel{!}{=} 0$$

so that

$$\widehat{v} = \frac{1}{n}\|\widetilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\theta}\|^2$$

$$\partial_v J(\widehat{\boldsymbol{\theta}}, v) = -\frac{1}{2v^2}\|\widetilde{\mathbf{y}} - \mathbf{X}\widehat{\boldsymbol{\theta}}\|^2 + \frac{n}{2v} \stackrel{!}{=} 0 \qquad \left(+ \frac{1}{2v^2}\|\widetilde{\mathbf{y}} - \mathbf{X}\widehat{\boldsymbol{\theta}}\|^2 \right)$$

$$\iff \frac{n}{2v} = \frac{1}{2v^2}\|\widetilde{\mathbf{y}} - \mathbf{X}\widehat{\boldsymbol{\theta}}\|^2 \qquad \left(\cdot \frac{2v^2}{n} \right)$$

$$\iff v = \frac{1}{n}\|\widetilde{\mathbf{y}} - \mathbf{X}\widehat{\boldsymbol{\theta}}\|^2$$

is a stationary point. The second derivative is positive, $\partial_v^2 J(\widehat{\boldsymbol{\theta}}, v) > 0$, at $\widehat{v}$ which is therefore a minimizer.

(b) Evaluating the learned parameter with the two data points, we have:

$$\widehat{\boldsymbol{\theta}} \approx \begin{bmatrix} -0.32 \\ 0.01 \end{bmatrix} \qquad \widehat{v} = 0$$

We obtain $\widehat{v} = 0$ since $\widehat{\boldsymbol{\theta}}$ fits the two data points perfectly. Thus, the norm is zero. If we use a $\widehat{\boldsymbol{\theta}}$ that is rounded to two digits after the decimal, we obtain $\widehat{v} \approx 0.16$.

so that the predicted incubation time is

$$f(\mathbf{x}; \widehat{\boldsymbol{\theta}}, \widehat{v}) = \exp(\mathbf{x}^T \widehat{\boldsymbol{\theta}} + 0) \approx 1.27$$

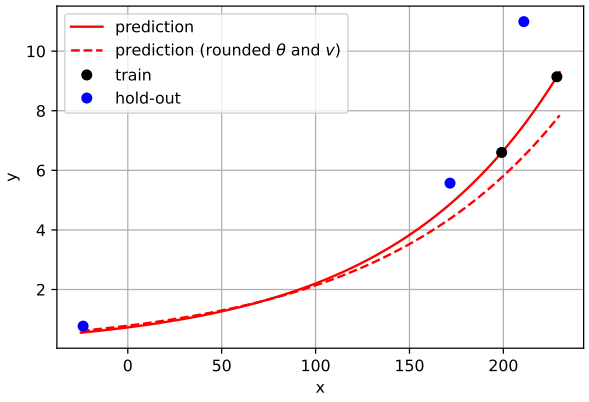
Note that this is a positive number.

We obtain 1.30 if we use $\widehat{\boldsymbol{\theta}}$ and $\widehat{v}$ that were rounded to two digits after the decimal.

(c) We have

$$E_{\text{hold-out}} = \frac{1}{3}\sum_{i=3}^{5}(y_i - f(\mathbf{x}; \widehat{\boldsymbol{\theta}}, \widehat{v}))^2 \approx 17.78$$

We obtain 4.19 if we use the computed $\widehat{\boldsymbol{\theta}}$ and $\widehat{v}$. If we round both to two digits after the decimal, we obtain 7.26. Rounding matters!



(d) While this figure is substantially lower than for the linear model, it is only a finite sample point estimate of the new error. To conclude that the new error is indeed smaller with confidence, we would need to quantify the uncertainty of the estimate.

4. (a) With the given model parameters $\hat{\theta}$, we calculate the probability that team A wins for each datapoint and add the results to our table:

i	x_1	x_2	$\hat{y}$	$p(y=1\|\mathbf{x}_i)$
1	55	5	1	≈ 0.95
2	90	1	-1	≈ 0.41
3	-20	0	-1	≈ 0.10
4	100	2	-1	≈ 0.65
5	-103	4	1	≈ 0.68

Now, we can find classification thresholds r with which the given predictions $\hat{y}$ can be reproduced for all datapoints in the table. This is the case as long as we choose a classification threshold on the interval $p(y=1|\mathbf{x}_4) < r < p(y=1|\mathbf{x}_5)$. In other words: $a = p(y=1|\mathbf{x}_4)$ and $b = p(y=1|\mathbf{x}_5)$

(b) Consider a datapoint ${\mathbf{x}_*}^T = [x_1, x_2]^T = [0,0]^T$. If the model has no intercept, i.e. $\theta_0 = 0$, then $p(y=1|\mathbf{x}_*) = 0.5$. In our case, $p(y=1|\mathbf{x}_*) = 0.12$, such that the probability that team B wins is significantly larger than the probability that team A wins. If x_1 and x_2 are the only features based on which the winning team is determined, a model without intercept seems more reasonable.

(c) Note, that $p(y < 2|\mathbf{x};\theta) = p(y=0|\mathbf{x};\theta) + p(y=1|\mathbf{x};\theta)$. Now, for each of these terms, the values only have to be put into the Poisson distribution with parameter

$$\lambda = \exp(\theta^T\mathbf{x}) = \exp([-17, 0.11, 2.35][1, 55, 5]^T) \approx 2.2$$
$$p(y=0|\mathbf{x},\theta) = \frac{\lambda^0}{0!}e^{-\lambda} \approx 0.108$$
$$p(y=1|\mathbf{x},\theta) = \frac{\lambda^1}{1!}e^{-\lambda} \approx 0.240$$

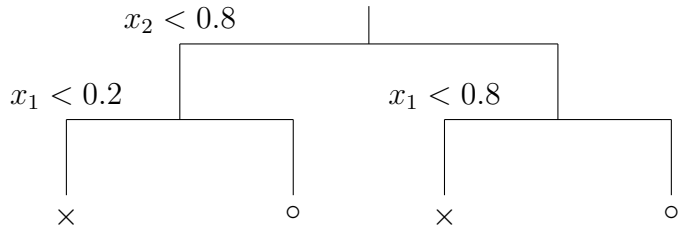
Thus, $p(y < 2|\mathbf{x};\theta) \approx 0.348$.

(d) The maximum likelihood starts by maximizing the likelihood of the dataset

$$\begin{aligned}
\arg\max_{\theta} \ell(\theta) &= \arg\max_{\theta} p(y_1, \dots y_n | \mathbf{x}_1, \dots \mathbf{x}_n; \theta) \\
&= \arg\max_{\theta} \prod_{i=1}^{n} p(y_i | \mathbf{x_i}; \theta) \\
&= \arg\max_{\theta} \log \left(\prod_{i=1}^{n} p(y_i | \mathbf{x_i}; \theta) \right) \\
&= \arg\max_{\theta} \sum_{i=1}^{n} \log(p(y_i | \mathbf{x_i}; \theta)) \\
&= \arg\max_{\theta} \sum_{i=1}^{n} \log \left(\frac{\lambda_i^{y_i}}{y_i!} e^{-\lambda_i} \right) \\
&= \arg\max_{\theta} \sum_{i=1}^{n} \log(\lambda_i^{y_i}) - \lambda_i - \log(y_i!) \\
&= \arg\max_{\theta} \sum_{i=1}^{n} y_i \theta^T \mathbf{x}_i - e^{\theta^T \mathbf{x}_i} - \log(y_i!) \\
&= \arg\max_{\theta} \sum_{i=1}^{n} y_i \theta^T \mathbf{x}_i - e^{\theta^T \mathbf{x}_i}
\end{aligned}$$

It is also possible to minimize the negative log likelihood equivalently. No closed form solution for this problem exists but instead numerical methods have to be used.

5. a) Sketch a binary classification tree with *depth=2*:



b) Increasing the depth of a decision tree would reduce the bias but increase the variance due to overfitting. Boosting is more suitable for shallow trees because it can effectively reduce the bias and thus learn a good model from high-bias base models (e.g. the decision stump).

c) **Mean:**

$$\mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B} z_b\right] = \sum_{b=1}^{B}\mathbb{E}[z_b] = \mu$$

Variance: Following hint 3, we first compute expression for $\mathbb{E}[z_i z_j]$ from correlation ρ:

$$\begin{aligned}\rho &= \frac{1}{\sigma^2}\mathbb{E}[(z_i-\mu)(z_j-\mu)], \quad i \neq j \\ \rho\sigma^2 &= \mathbb{E}[z_i z_j - z_i\mu - z_j\mu + \mu^2] \\ \mathbb{E}[z_i z_j] &= \rho\sigma^2 + \mu^2\end{aligned}$$

since $\mathbb{E}[z_i]\mu = \mu^2$. With the known equation for the second raw statistical moment, we can write

$$\mathbb{E}[z_i z_j] = \begin{cases}\sigma^2 + \mu^2 & i = j, \\ \rho\sigma^2 + \mu^2 & i \neq j\end{cases}$$

Now we can use hint 1 to write:

$$\begin{aligned}\mathrm{Var}\left[\frac{1}{B}\sum_{b=1}^{B} z_b\right] &= \frac{1}{B^2}\mathbb{E}\left[(\sum_{b=1}^{B} z_b)^2\right] \\ &= \frac{1}{B^2}\left(\mathbb{E}\left[(\sum_{b=1}^{B} z_b)^2\right] - \mathbb{E}\left[\sum_{b=1}^{B} z_b\right]^2\right)\end{aligned}$$

The second term is simple to solve:

$$\mathbb{E}\left[\sum_{b=1}^{B} z_b\right]^2 = (\sum_{b=1}^{B}\mathbb{E}\left[z_b\right])^2 = B^2\mu^2$$

Then, let us solve the first term, where we use hint 2:

$$\begin{aligned}
\mathbb{E}\left[(\sum_{b=1}^{B} z_b)^2\right] &= \mathbb{E}\left[\sum_{i,j=1}^{B} z_i z_j\right] \\
&= \sum_{i,j=1}^{B} \mathbb{E}\left[z_i z_j\right] \\
&= B\left(\mathbb{E}[z_i z_i] + (B-1)\mathbb{E}[z_i z_j]\right), \quad i \neq j \\
&= B(\sigma^2 + \mu^2) + (B^2 - B)(\rho\sigma^2 + \mu^2)
\end{aligned}$$

where in the last row we used our solution from hint 3. Now we can combine the two terms for the variance to obtain:

$$\begin{aligned}
\text{Var}\left[\frac{1}{B}\sum_{b=1}^{B} z_b\right] &= \frac{1}{B}(\sigma^2 + \mu^2) + (1 - \frac{1}{B})(\rho\sigma^2 + \mu^2) - \mu^2 \\
&= \frac{1-\rho}{B}\sigma^2 + \rho\sigma^2
\end{aligned}$$

This is the solution for the variance as also given in the formula sheet (see "Sum of identically distributed variables:").

d) Given the equation

$$\text{Var}\left[\frac{1}{B}\sum_{b=1}^{B} z_b\right] = \frac{1-\rho}{B}\sigma^2 + \rho\sigma^2$$

from the formula sheet, we directly observe that increasing B reduces the first term as long. However, this only holds if $\rho < 1$ as otherwise the first term is eliminated. Generally, we can assume that $\rho < 1$ for Bagging.