UNIVERSITY OF OXFORD

Machine Learning
Michaelmas Term 2021
Week 5

Problem Sheet 2 Solutions

1 Maximum Likelihood Estimation of σ

As presented in Lecture 4, we consider a discriminative framework, where the input datapoints $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are fixed (we will not consider these as being generated by a random process). Let $\mathbf{w}$ and σ be the parameters defining the linear model with Gaussian noise, *i.e.,*

$$y_i \sim \mathcal{N}(\mathbf{x}_i^\mathsf{T}\mathbf{w}, \sigma^2). \tag{1.1}$$

In Lecture 4 we showed that the maximum likelihood estimate for $\mathbf{w}$ is the same as the least square estimator, $\mathbf{w}_{\text{ML}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$. Show that the MLE for σ^2 is given by

$$\sigma^2_{\text{ML}} = \frac{1}{N}(\mathbf{y} - \mathbf{X}\mathbf{w}_{\text{ML}})^\mathsf{T}(\mathbf{y} - \mathbf{X}\mathbf{w}_{\text{ML}}). \tag{1.2}$$

Solution: We begin by recalling the negative log-likelihood (NLL)

$$\text{NLL}(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma) = \frac{1}{2\sigma^2}(\mathbf{X}\mathbf{w} - \mathbf{y})^\mathsf{T}(\mathbf{X}\mathbf{w} - \mathbf{y}) + \frac{N}{2}\log(2\pi\sigma^2)$$

Differentiating with respect to σ

$$\frac{\mathrm{d}\text{NLL}(\mathbf{y} \mid \mathbf{X}, \mathbf{w}, \sigma)}{\mathrm{d}\sigma} = -\frac{1}{\sigma^3}(\mathbf{X}\mathbf{w} - \mathbf{y})^\mathsf{T}(\mathbf{X}\mathbf{w} - \mathbf{y}) + \frac{N}{\sigma}$$

Calculating for σ and substituting $\mathbf{w} = \mathbf{w}_{\text{ML}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$ gives the required result.

2 Centering and Ridge Regression

Assume that $\frac{1}{N}\sum_{i=1}^N \mathbf{x}_i = \mathbf{0}$, *i.e.,* the data is centered. (In this question we will treat the constant term separately, as centering this would give us a column of 0s.) Let us denote the parameter for the leading constant term as b (for "bias"). So the linear model is $\widehat{y} = b + \mathbf{x}^\mathsf{T}\mathbf{w}$. Consider minimizing the ridge objective:

$$\mathcal{L}_{\text{ridge}}(\mathbf{w}, b) = (\mathbf{X}\mathbf{w} + b\mathbf{1} - \mathbf{y})^\mathsf{T}(\mathbf{X}\mathbf{w} + b\mathbf{1} - \mathbf{y}) + \lambda\mathbf{w}^\mathsf{T}\mathbf{w} \tag{2.1}$$

Here $\mathbf{1}$ is the vector of all ones and note that b^2 is not regularized. Show that if $\widehat{b}$ and $\widehat{\mathbf{w}}$ are the resulting solutions obtained by minimising the above objective, then

$$\widehat{b} = \frac{1}{N}\sum_{i=1}^N y_i$$
$$\widehat{\mathbf{w}} = (\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I}_D)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$$

What happens if you also center $\mathbf{y}$?

Solution: The fact that $\mathbf{X}$ is centered implies that $\mathbf{X}^\mathsf{T}\mathbf{1} = \mathbf{0}$. When $\mathbf{y}$ is centered, $\mathbf{y}^\mathsf{T}\mathbf{1} = 0$. We use these facts in the computation of the derivative below. We calculate the derivative of $\mathcal{L}_{\text{ridge}}(\mathbf{w}, b)$ with respect to b and the gradient with respect to $\mathbf{w}$.

$$\frac{\mathrm{d}\mathcal{L}_{\text{ridge}}}{\mathrm{d}b}(\mathbf{w}, b) = 2Nb - 2 \cdot \mathbf{1}^\mathsf{T}\mathbf{y}$$
$$\nabla_{\mathbf{w}}\mathcal{L}_{\text{ridge}}(\mathbf{w}, b) = 2\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w} - 2\mathbf{X}^\mathsf{T}\mathbf{y} + 2\lambda\mathbf{w}$$

When only $\mathbf{X}$ is centered, we get that $\widehat{b} = \frac{1}{N}\sum_{i=1}^{N} y_i$ by setting the derivative to 0. When $\mathbf{y}$ is centered, $b = 0$, so we can ignore the bias term entirely.

3 Bias of the Least Squared Estimator

Suppose that the data $\mathcal{D} = \langle(\mathbf{x}_i, y_i)\rangle_{i=1}^{N}$ is truly generated from a linear model, *i.e.,*

$$\mathbb{E}\left[y \mid \mathbf{x}, \mathbf{w}^*\right] = \mathbf{x}^\mathsf{T}\mathbf{w}^* \tag{3.1}$$

for some fixed (but unknown) parameter vector $\mathbf{w}^*$. Recall that the least squares estimator is

$$\widehat{\mathbf{w}}_{\text{LS}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}. \tag{3.2}$$

1. Assume that $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are fixed and that $\mathbf{X}^\mathsf{T}\mathbf{X}$ is invertible. You can think of the data $\mathcal{D}$ as a random variable (because of the possible noise in the y_is). Thus, $\widehat{\mathbf{w}}_{\text{LS}}(\mathcal{D})$ is itself a random variable. Show that the expectation of the estimator $\widehat{\mathbf{w}}_{\text{LS}}(\mathcal{D})$ (with respect to $\mathcal{D}$) is $\mathbf{w}^*$. Such an estimator is called an *unbiased* estimator, as its expectation equals the true parameter value.

   Solution: We observe that $\mathbb{E}\left[\mathbf{y} \mid \mathbf{X}, \mathbf{w}^*\right] = \mathbf{X}\mathbf{w}^*$. Thus,

   $$\mathbb{E}\left[\widehat{\mathbf{w}}_{\text{LS}}\right] = \mathbb{E}\left[(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}\right] = \mathbf{w}^*.$$

2. Now suppose we have some other estimator $\widehat{\mathbf{w}}$ which may not be unbiased. The bias is defined as

   $$\text{Bias}(\widehat{\mathbf{w}}) = \|\mathop{\mathbb{E}}_{\mathcal{D}}\left[\widehat{\mathbf{w}}(\mathcal{D})\right] - \mathbf{w}^*\|. \tag{3.3}$$

   Thus, the bias is the Euclidean distance between the expectation of the estimator and the true parameter. Suppose you are interested in minimizing the squared distance between the estimated parameters and true parameters, *i.e.,* to minimize $\|\widehat{\mathbf{w}}(\mathcal{D}) - \mathbf{w}^*\|^2$. Show that the expected (with respect to $\mathcal{D}$) squared distance can be decomposed as follows:

   $$\mathop{\mathbb{E}}_{\mathcal{D}}\left[\|\widehat{\mathbf{w}}(\mathcal{D}) - \mathbf{w}^*\|^2\right] = \|\mathop{\mathbb{E}}_{\mathcal{D}}\left[\widehat{\mathbf{w}}(\mathcal{D})\right] - \mathbf{w}^*\|^2 + \mathop{\mathbb{E}}_{\mathcal{D}}\left[\|\widehat{\mathbf{w}}(D) - \mathop{\mathbb{E}}_{\mathcal{D}}\left[\widehat{\mathbf{w}}(\mathcal{D})\right]\|^2\right] \tag{3.4}$$

   The first term above is just the squared bias and the second term above is the variance of the estimator. Thus, while being unbiased looks like a natural property to demand of estimators, it might sometimes be preferable to have a *biased* estimator if it has a much

lower variance. This is what ridge regression or LASSO does.

Solution: We have the following.

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{D}}\left[\|\widehat{\mathbf{w}}(\mathcal{D}) - \mathbf{w}^*\|^2\right] &= \mathop{\mathbb{E}}_{\mathcal{D}}\left[\|\widehat{\mathbf{w}}(\mathcal{D}) - \mathop{\mathbb{E}}_{\mathcal{D}}\left[\widehat{\mathbf{w}}(\mathcal{D})\right] + \mathop{\mathbb{E}}_{\mathcal{D}}\left[\widehat{\mathbf{w}}(\mathcal{D})\right] - \mathbf{w}^*\|^2\right] \\
&= \mathop{\mathbb{E}}_{\mathcal{D}}\left[\|\widehat{\mathbf{w}}(\mathcal{D}) - \mathop{\mathbb{E}}_{\mathcal{D}}\left[\widehat{\mathbf{w}}(\mathcal{D})\right]\|^2\right] + \|\mathop{\mathbb{E}}_{\mathcal{D}}\left[\widehat{\mathbf{w}}(\mathcal{D})\right] - \mathbf{w}^*\|^2 \\
&\quad + 2\left(\mathop{\mathbb{E}}_{\mathcal{D}}\left[\widehat{\mathbf{w}}(\mathcal{D})\right] - \mathbf{w}^*\right)\mathbb{E}_{\mathcal{D}}\left[\widehat{\mathbf{w}}(\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[\widehat{\mathbf{w}}(\mathcal{D})]\right] \\
&= \|\mathbb{E}_{\mathcal{D}}[\widehat{\mathbf{w}}(\mathcal{D})] - \mathbf{w}^*\|^2 + \mathbb{E}_{\mathcal{D}}\left[\|\widehat{\mathbf{w}}(\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[\widehat{\mathbf{w}}(\mathcal{D})]\|^2\right]
\end{aligned}$$

4 Maximum Likelihood and Model Selection

Let the random variable $x \in \{0, 1\}$ model the outcome of an experiment, such that the event $x = 1$ occurs with probability θ_1. Suppose that someone else observes the experiment and reports to you the outcome, y. But this person is unreliable and only reports the result correctly with probability θ_2. That is, $p(y \mid x, \theta_2)$ is given by

	$y = 0$	$y = 1$
$x = 0$	θ_2	$1 - \theta_2$
$x = 1$	$1 - \theta_2$	θ_2

Assume that θ_2 is independent of x and θ_1.

1. Write down the joint probability distribution $p(x, y \mid \boldsymbol{\theta})$ as a 2×2 table, in terms of $\boldsymbol{\theta} = (\theta_1, \theta_2)$.

   Solution: The joint distribution is $p(x, y \mid \boldsymbol{\theta}) = p(x \mid \theta_1)p(y \mid x, \theta_2)$, given by the following table:

	$y = 0$	$y = 1$
$x = 0$	$(1 - \theta_1)\theta_2$	$(1 - \theta_1)(1 - \theta_2)$
$x = 1$	$\theta_1(1 - \theta_2)$	$\theta_1\theta_2$

2. Given the following dataset: $\mathbf{x} = (1, 1, 0, 1, 1, 0, 0)$, $\mathbf{y} = (1, 0, 0, 0, 1, 0, 1)$. What are the numerical values of the MLEs for θ_1 and θ_2? What is the numerical value $p(\mathcal{D} \mid \widehat{\boldsymbol{\theta}}, M_2)$ where M_2 denotes this 2-parameter model? Justify your answer by including the derivations.

   Solution: The log-likelihood is

$$\log p(\mathcal{D} \mid \boldsymbol{\theta}) = \sum_i \log p(x_i \mid \theta_1) + \sum_i \log p(y_i \mid x_i, \theta_2)$$

   Hence we can optimize each term separately. For θ_1, we have

$$\widehat{\theta}_1 = \frac{\sum_i \mathbb{1}(x_i = 1)}{N} = \frac{|\{i : x_i = 1\}|}{N} = \frac{4}{7} = 0.5714$$

For θ_2, we have

$$\widehat{\theta}_2 = \frac{\sum_i \mathbb{1}(x_i = y_i)}{N} = \frac{|\{i : x_i = y_i\}|}{N} = \frac{4}{7} = 0.5714$$

Let $N(x = 1)$ and $N(x = 0)$ be the number of x observations with values 1 and 0 respectively; let $N(x = y)$ and $N(x \neq y)$ be the number of observations with equal and unequal values for x and y respectively. The likelihood is

$$\begin{aligned} p(\mathcal{D} \mid \widehat{\boldsymbol{\theta}}, M_2) &= (\frac{4}{7})^{N(x=1)}(\frac{3}{7})^{N(x=0)}(\frac{4}{7})^{N(x=y)}(\frac{3}{7})^{N(x\neq y)} \\ &= (\frac{4}{7})^4(\frac{3}{7})^3(\frac{4}{7})^4(\frac{3}{7})^3 \\ &= (\frac{4}{7})^8(\frac{3}{7})^6 \approx 7.04 \times 10^{-5} \end{aligned}$$

3. Now consider a model with 4 parameters, $\boldsymbol{\theta} = (\theta_{0,0}, \theta_{0,1}, \theta_{1,0}, \theta_{1,1})$, representing $p(x, y \mid \boldsymbol{\theta}) = \theta_{x,y}$. (Only 3 of these parameters are free to vary, since they must sum to one.) What is the MLE of $\boldsymbol{\theta}$? What is $p(\mathcal{D} \mid \widehat{\boldsymbol{\theta}}, M_4)$ where M_4 denotes this 4-parameter model?

   Solution: The table of joint counts is

	$y = 0$	$y = 1$
$x = 0$	2	1
$x = 1$	2	2

   We can think of this as a multinomial distribution with 4 states. Normalizing the counts gives the MLE:

	$y = 0$	$y = 1$
$x = 0$	2/7	1/7
$x = 1$	2/7	2/7

   Let $N(x = b_1, y = b_2)$ denote the number of observations with x value b_1 and y value b_2 for $b_1, b_2 \in \{0, 1\}$. The likelihood is

$$\begin{aligned} p(\mathcal{D} \mid \widehat{\boldsymbol{\theta}}, M_4) &= \theta_{00}^{N(x=0,y=0)}\theta_{01}^{N(x=0,y=1)}\theta_{10}^{N(x=1,y=0)}\theta_{11}^{N(x=1,y=1)} \\ &= (\frac{2}{7})^2(\frac{1}{7})^1(\frac{2}{7})^2(\frac{2}{7})^2 \\ &= (\frac{2}{7})^6(\frac{1}{7})^1 \approx 7.77 \times 10^{-5} \end{aligned}$$

   This likelihood is higher than the previous likelihood, because the model has more parameters.

4. Suppose we are not sure which model is correct. We compute the leave-one-out cross-validated log-likelihood of the 2-parameter model and the 4-parameter model as follows:

$$L(M) = \sum_{i=1}^{N} \log p(x_i, y_i \mid M, \widehat{\theta}(\mathcal{D}_{-i}))$$

and $\widehat{\theta}(\mathcal{D}_{-i})$ denotes the MLE computed on $\mathcal{D}$ excluding the i^{th} observation. Which model will CV pick and why?

Solution: For M_4, when we omit case 7, we will have $\widehat{\theta}_{01} = 0$, so $p(x_7, y_7 \mid M_4, \widehat{\theta}) = 0$, so $L(M_4) = -\infty$. However, $L(M_2)$ will be finite, since all counts remain non zero when we leave out a single case.

Hence LOOCV will prefer M_2, since M_4 is overfitting.

5 The *Huber loss* in a linear regression setting

In this question, we will investigate the *Huber loss* in a linear regression setting. Given arbitrary but fixed parameters $\lambda, \mu \in \mathbb{R}$ such that $\lambda, \mu > 0$, the Huber loss is given by the function $h_{\lambda,\mu} : \mathbb{R} \to \mathbb{R}$ such that

$$h_{\lambda,\mu}(z) = \begin{cases} \lambda\left(|z| - \frac{\lambda}{4\mu}\right) & \text{if } |z| \geq \frac{\lambda}{2\mu}, \\ \mu z^2 & \text{otherwise.} \end{cases}$$

Given a vector $\mathbf{z} = (z_1, \ldots, z_D) \in \mathbb{R}^D$, we extend $h_{\lambda,\mu}$ such that $h_{\lambda,\mu}(\mathbf{z}) = \sum_{i=1}^{D} h_{\lambda,\mu}(z_i)$. Recall that when dealing with absolute values, the sign function defined as follows is often helpful:

$$\text{sign}(z) = \begin{cases} 1 & \text{if } z > 0, \\ -1 & \text{otherwise.} \end{cases}$$

1. Let us fix $\lambda = 4$ and $\mu = 1$. Draw three graphs plotting $h_{4,1}(z)$, and the absolute and the square loss functions. Briefly compare the Huber loss to the absolute and the square loss functions. What can you say about outliers?

   Solution: The absolute loss function is given by $f(z) = |z|$, and the square loss by $g(z) = z^2$. The three functions are depicted in Figure 1. For $|z| \geq 2$, $h_{4,1}$ penalises exactly as the absolute loss, and for $|z| < 2$ in exactly the same way as the square loss. Just like the absolute loss, the Huber loss will not be as sensitive to extreme outliers as the square loss, but still impose a quadratic penality for all z such that $|z| < 2$. Thus, the Huber loss can be seen as a combination of the absolute and square loss.

2. Given a training set $\mathcal{D} = \langle(\mathbf{x}_i, y_i)\rangle_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^D$ and $y_i \in \mathbb{R}$, we ignore any noise terms and define the loss function as

   $$\mathcal{L}(\mathbf{w}; \mathcal{D}) = \frac{1}{N} \sum_{i=1}^{N} h_{\lambda,\mu}(y_i - \mathbf{w}^\intercal \cdot \mathbf{x}_i),$$

   where $\mathbf{w} \in \mathbb{R}^D$ is the model parameter.

   Compute $\nabla_{\mathbf{w}} \mathcal{L}$.

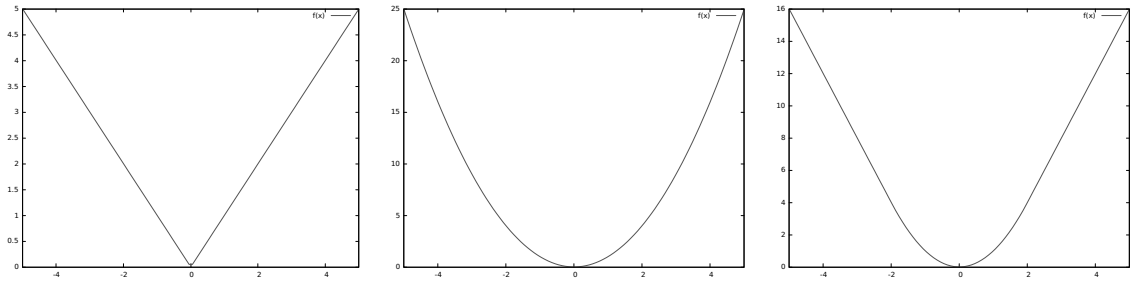

Figure 1: From left to right: absolute loss, square loss, and Huber loss with $\lambda = 4$ and $\mu = 1$.

Solution: Observe that $h_{\lambda,\mu}$ is defined piece-wise by the functions $f(z) = \lambda(|z| - \frac{\lambda}{4\mu})$ and $g(z) = \mu z^2$. We have

$$f'(z) = \text{sign}(z)\lambda \qquad \text{and} \qquad g'(z) = 2\mu z.$$

Hence,

$$h'_{\lambda,\mu}(z) = \begin{cases} \text{sign}(z)\lambda & \text{if } |z| \geq \frac{\lambda}{2\mu} \\ 2\mu z & \text{otherwise} \end{cases}.$$

Let $z_i(\mathbf{w}) = y_i - \mathbf{w}^\intercal \cdot \mathbf{x}_i$, by the chain rule we have

$$\frac{\partial h_{\lambda,\mu}(z_i(\mathbf{w}))}{\partial w_j} = -h'_{\lambda,\mu}(z_i(\mathbf{w}))\mathbf{x}_{i,j},$$

where $\mathbf{x}_{i,j}$ denotes the j-th component of $\mathbf{x}_i$. Consequently,

$$\nabla_{\mathbf{w}} h_{\lambda,\mu}(z_i(\mathbf{w})) = -h'_{\lambda,\mu}(z_i(\mathbf{w}))\mathbf{x}_i.$$

We conclude

$$\begin{aligned} \nabla_{\mathbf{w}}\mathcal{L} &= \frac{1}{N}\sum_{i=1}^{N} \nabla_{\mathbf{w}} h_{\lambda,\mu}(y_i - \mathbf{w}^\intercal \cdot \mathbf{x}_i) \\ &= -\frac{1}{N}\sum_{i=1}^{N} h'_{\lambda,\mu}(y_i - \mathbf{w}^\intercal \cdot \mathbf{x}_i) \cdot \mathbf{x}_i. \end{aligned}$$

3. For the remainder of this question, we will discuss using the Huber loss as a regulariser. Let $\ell : \mathbb{R} \to \mathbb{R}$ be an arbitrary loss function, and consider the following regularised loss functions:

$$\mathcal{H}(\mathbf{z}; \mathcal{D}) = h_{\lambda,\mu}(\mathbf{z}) + \frac{1}{N}\sum_{i=1}^{N} \ell(y_i - \mathbf{z}^\intercal \cdot \mathbf{x}_i),$$

$$\mathcal{S}(\mathbf{v}, \mathbf{w}; \mathcal{D}) = \lambda||\mathbf{v}||_1 + \mu||\mathbf{w}||_2^2 + \frac{1}{N}\sum_{i=1}^{N} \ell(y_i - (\mathbf{v} + \mathbf{w})^\intercal \cdot \mathbf{x}_i).$$

Suppose we let $\lambda \to \infty$ in $\mathcal{H}(\mathbf{z}; \mathcal{D})$ and $\mathcal{S}(\mathbf{v}, \mathbf{w}; \mathcal{D})$. Which types of regularised regression do we obtain? What happens when $\mu \to \infty$?

Solution: As $\lambda \to \infty$, we have $||\mathbf{v}||_1 \to 0$ for an optimal solution of $\mathcal{S}(\mathbf{v}, \mathbf{w}; \mathcal{D})$. Likewise, as $\lambda \to \infty$ we have that $h_{\lambda,\mu}(z)$ converges to μz^2, and hence $h_{\lambda,\mu}(\mathbf{z}) \to \mu||\mathbf{z}||_2^2$. The resulting regularised regression problems are called ridge regression.

Similarly, as $\mu \to \infty$, we have $||\mathbf{w}||_2^2 \to 0$ for an optimal solution of $\mathcal{S}(\mathbf{v}, \mathbf{w}; \mathcal{D})$. Symmetrically, $h_{\lambda,\mu}(\mathbf{z}) \to \lambda||\mathbf{z}||_1$. The resulting regularised regression problem is called Lasso.