Machine Learning	Prof. Matthias Hein
Exercise Sheet 2 - 6.5.2021 (12 points)	due: 13.5.2021

Exercise 4 - Maximum Likelihood and Maximum A Posteriori Estimation

- **(3 Points)** Consider the regression problem where the input $X \in \mathbb{R}^d$ and the output $Y \in \mathbb{R}$. Assume that the likelihood is specified in terms of the unknown parameter $w \in \mathbb{R}^d$ as

$$p(y|x,w) = \frac{1}{\sqrt{2\pi g(x)}} e^{-\frac{(y-\langle w,x\rangle)^2}{2g(x)}},$$

where $g : \mathbb{R}^d \to \mathbb{R}_+^*$ is a known positive function. Compute the maximum likelihood estimator of w.

- **(3 points)** Further we have a prior distribution on w:

$$p(w) = \frac{1}{(2\pi)^{\frac{d}{2}} (\prod_{i=1}^d \lambda_i)^{\frac{1}{2}}} e^{-\frac{1}{2}\langle w, \Lambda^{-1} w\rangle}.$$

where Λ is a diagonal matrix with the diagonal entries given by $\lambda_i > 0$ for all $i \in \{1, ..., d\}$. We are given an i.i.d. training sample $(x_i, y_i)_{i=1}^n$, which we assume to be additionally conditionally independent given the model. We impose the condition that w is independent of x. Use this first to show that

$$p(y, x \,|\, w) = p(y \,|\, w, x) p(x).$$

What is the maximum a posteriori (MAP) estimator of w ?

Points split:

- derive ML estimator (3)
- show the identity holds (1)
- derive MAP estimator (2)

Solution

a. The likelihood function is defined as

$$\mathcal{L}_n(w) = \prod_{i=1}^n p(y_i|w, x_i),$$

where $(x_i, y_i)_{i=1}^n$ is the i.i.d. training sample.

Using the fact that ln is a strictly increasing function, we have

$$\arg\max_w \prod_{i=1}^n p(y_i|w, x_i) = \arg\max_w \sum_{i=1}^n \ln p(y_i|w, x_i) = \arg\max_w \sum_{i=1}^n - \ln \sqrt{2\pi g(x_i)} - \frac{(y_i - \langle w, x_i\rangle)^2}{2g(x_i)}.$$

Since the first term is constant w.r.t. the optimization, it can be ignored. Thus we have

$$\arg\max_w \sum_{i=1}^n -\frac{(y_i - \langle w, x_i\rangle)^2}{2g(x_i)} = \arg\min_w \sum_{i=1}^n \frac{(y_i - \langle w, x_i\rangle)^2}{2g(x_i)} =: \Psi(w)$$

We see that each term in the above sum is a convex function in the variable w, noting that $g(x_i)$, constant w.r.t. w, are positive. Letting Γ be the diagonal matrix with elements $\Gamma_{ii} = \frac{1}{g(x_i)}$, the objective $\Psi(w)$ can be written as

$$\Psi(w) = \frac{1}{2} \langle Y - Xw, \Gamma(Y - Xw) \rangle .$$

Thus a necessary and sufficient condition for a minimizer is

$$\begin{aligned}(\nabla_w \Psi)(w) = 0 &\implies X^T \Gamma X w - X^T \Gamma Y = 0 \\ &\implies w = (X^T \Gamma X)^{-1} X^T \Gamma Y.\end{aligned}$$

b. NOTE: in contrast to the original formulation we require additionally that the sample $(X_i, Y_i)_{i=1}^n$ is conditionally independent given the model. Otherwise one of the steps below cannot be done if just assume that the sample is i.i.d.

We have

$$p(y, x \,|\, w) = \frac{p(x, y, w)}{p(w)} = \frac{p(y \,|\, w, x) p(x, w)}{p(w)} = p(y \,|\, w, x) p(x \,|\, w) = p(y \,|\, w, x) p(x),$$

where the last step follows from the independence of w and x.

The posterior of w given the i.i.d. training sample is

$$p\Big(w | (x_i, y_i)_{i=1}^n\Big) = \frac{p\Big((x_i, y_i)_{i=1}^n | w\Big)\ p(w)}{p\Big((x_i, y_i)_{i=1}^n\Big)}.$$

The maximum a posteriori estimate is then the "mode" (maximizer) of the posterior. Thus we have,

$$\begin{aligned}
w_{\text{MAP}} = \operatorname*{arg\,max}_{w} \frac{p\Big((x_i, y_i)_{i=1}^n | w\Big)\ p(w)}{p\Big((x_i, y_i)_{i=1}^n\Big)} &= \operatorname*{arg\,max}_{w \in \mathbb{R}^d} p\Big((x_i, y_i)_{i=1}^n | w\Big)\ p(w) \\
&= \operatorname*{arg\,max}_{w \in \mathbb{R}^d} \prod_{i=1}^n p\Big((x_i, y_i) | w\Big)\ p(w) \qquad \text{(samples } (x_i, y_i)_{i=1}^n \text{ are conditionally independent given } w) \\
&= \operatorname*{arg\,max}_{w \in \mathbb{R}^d} \prod_{i=1}^n p(y_i | w, x_i)\ p(w)\ p(x_i) \qquad \text{(above calculation using independence of } w \text{ and } x) \\
&= \operatorname*{arg\,max}_{w \in \mathbb{R}^d} \prod_{i=1}^n p(y_i | w, x_i)\ p(w) \qquad \Big(p(x_i) \text{ is a constant w.r.t. w}\Big) \\
&= \operatorname*{arg\,max}_{w \in \mathbb{R}^d} \sum_{i=1}^n \ln p(y_i | w, x_i)\ + \ln p(w) \qquad (\ln \text{ is a strictly increasing function}) \\
&= \operatorname*{arg\,max}_{w \in \mathbb{R}^d} \sum_{i=1}^n - \frac{(y_i - \langle w, x_i \rangle)^2}{2 g(x_i)} - \frac{1}{2} \left\langle w, \Lambda^{-1} w \right\rangle \qquad \text{(constant terms do not affect the minimizer)} \\
&= \operatorname*{arg\,min}_{w \in \mathbb{R}^d} \sum_{i=1}^n \frac{(y_i - \langle w, x_i \rangle)^2}{g(x_i)} + \left\langle w, \Lambda^{-1} w \right\rangle \qquad \text{(switching the sign of the objective)} \\
&= \operatorname*{arg\,min}_{w \in \mathbb{R}^d} \sum_{i=1}^n \gamma_i (y_i - \langle w, x_i \rangle)^2 + \left\langle w, \Lambda^{-1} w \right\rangle =: \Psi(w). \qquad \Big(\text{letting } \gamma_i = \frac{1}{g(x_i)}\Big)
\end{aligned}$$

The objective $\Psi(w)$ can be written in matrix-vector notation as

$$\Psi(w) = \langle Y - Xw, \Gamma(Y - Xw) \rangle + \left\langle w, \Lambda^{-1} w \right\rangle ,$$

where $X \in \mathbb{R}^{n \times d}$ is the design matrix whose rows contain the inputs x_i of the training sample, $Y \in \mathbb{R}^n$ is the vector containing the outputs y_i and Γ is the diagonal matrix whose i^{th} diagonal entry is given by γ_i.

Since $\Psi(w)$ is convex, a necessary and sufficient condition for a minimizer is

$$\begin{aligned}(\nabla_w \Psi)(w_{MAP}) = 0 &\implies 2X^T\Gamma X w_{MAP} - 2X^T\Gamma Y + 2\Lambda^{-1}w_{MAP} = 0\\ &\implies w_{MAP} = (X^T\Gamma X + \Lambda^{-1})^{-1}X^T\Gamma Y.\end{aligned}$$

Note that the inhomogeneous noise model (variance is equal to $g(x)$) corresponds to using a weighted loss, where the weight is inverse to the variance. This makes intuitively sense as points x for which the model is pretty sure (small variance $g(x)$) should be fitted very accurately, whereas for points x where the variance is large we should not penalize the errors too much,

Exercise 5 - ML and MAP estimators

Consider the two r.v., representing two sensors estimating the same value θ,

$$\begin{aligned}A =&\theta + \epsilon_1, \qquad \epsilon_1 \sim \mathcal{N}(0, \sigma_1^2)\\ B =&\theta + \epsilon_2, \qquad \epsilon_2 \sim \mathcal{N}(0, \sigma_2^2),\end{aligned}$$

with ϵ_1 and ϵ_2 independent, and their realizations $(a_1, \ldots, a_n)$ and $(b_1, \ldots, b_n)$ (in practice $n = 1$ and we need an estimation of θ).

- **(3 points)** Compute the MLE of θ, using the information from both sensors and assuming σ_1, σ_2 known.
- **(3 points)** If additionally we assume a prior

  $$p(\theta) = \mathcal{N}(\mu_P, \sigma_P^2),$$

  which is the MAP estimator of θ? With which σ_P would we get $\hat{\theta}_{ML} = \hat{\theta}_{MAP}$?

Points split:

- derive ML estimator (3)
- derive MAP estimator (2)
- derive the condition for $\hat{\theta}_{ML} = \hat{\theta}_{MAP}$ (1)

Solutions:

- We first notice that, given θ, $A \sim \mathcal{N}(\theta, \sigma_1^2)$ and $B \sim \mathcal{N}(\theta, \sigma_2^2)$. Moreover, we have that A and B are independent given θ since the noise ϵ_1 and ϵ_2 are independent. Then, we have

  $$p(a, b|\theta) = p(a|\theta) \cdot p(b|\theta) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(a-\theta)^2}{2\sigma_1^2}} \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(b-\theta)^2}{2\sigma_2^2}}$$

  and we get the likelihood

  $$\begin{aligned}\log L(\theta) = \log \Pi_i p(a_i|\theta)p(b_i|\theta) =& \sum_{i=1}^n \log\left(\frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(a_i-\theta)^2}{2\sigma_1^2}} \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(b_i-\theta)^2}{2\sigma_2^2}}\right)\\ =& \sum_{i=1}^n - \log(2\pi\sigma_1\sigma_2) - \frac{(a_i - \theta)^2}{2\sigma_1^2} - \frac{(b_i - \theta)^2}{2\sigma_2^2}.\end{aligned}$$

We can see that $-\log L(\theta)$ is convex in θ, so if we find an unique critical point, it is a global maximizer of L. Thus,

$$\frac{d}{d\theta}\log L(\theta) = \sum_{i=1}^{n}\left(\frac{a_i-\theta}{\sigma_1^2}+\frac{b_i-\theta}{\sigma_2^2}\right) = \sum_{i=1}^{n}\frac{a_i}{\sigma_1^2}+\frac{b_i}{\sigma_2^2} - n\theta\left(\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}\right) = 0$$

and

$$\hat{\theta}_{ML} = \left(\sum_{i=1}^{n}\frac{a_i}{\sigma_1^2}+\frac{b_i}{\sigma_2^2}\right) n^{-1}\left(\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}\right)^{-1}.$$

- The MAP estimator is given by maximizing

$$p(\theta|a_1,\dots,a_n,b_1,\dots,b_n) = p(a_1,\dots,a_n,b_1,\dots,b_n|\theta)p(\theta)/p(a_1,\dots,a_n,b_1,\dots,b_n).$$

Since $p(a_1,\dots,a_n,b_1,\dots,b_n)$ does not depend on θ and using the conditional independence of A and B as before, we can maximize

$$\begin{aligned}\log p(a_1,\dots,a_n,b_1,\dots,b_n|\theta)+\log p(\theta) &= \sum_{i=1}^{n}\log p(a_i|\theta)+\log p(b_i|\theta)+\log p(\theta)\\ &= \sum_{i=1}^{n}\left(-\log(2\pi\sigma_1\sigma_2)-\frac{(a_i-\theta)^2}{2\sigma_1^2}-\frac{(b_i-\theta)^2}{2\sigma_2^2}\right)-\log(\sqrt{2\pi}\sigma_P)-\frac{(\theta-\mu_P)^2}{2\sigma_P^2}\end{aligned}$$

whose critical points solve

$$\sum_{i=1}^{n}\left(\frac{a_i}{\sigma_1^2}+\frac{b_i}{\sigma_2^2}\right)-n\theta\left(\frac{1}{\sigma_1^2}+\frac{1}{\sigma_2^2}\right)-\frac{\theta}{\sigma_P^2}+\frac{\mu_P}{\sigma_P^2}=0.$$

Finally, we get

$$\hat{\theta}_{MAP} = \left(\frac{\sum_{i=1}^{n}a_i}{\sigma_1^2}+\frac{\sum_{i=1}^{n}b_i}{\sigma_2^2}+\frac{\mu_P}{\sigma_P^2}\right)\left(\frac{n}{\sigma_1^2}+\frac{n}{\sigma_2^2}+\frac{1}{\sigma_P^2}\right)^{-1}.$$

Note that for $\sigma_P^2 \to +\infty$ we find the ML estimator, which is the case when the prior becomes non informative (the variance of the normal distribution grows arbitrarily).