# Exercise Sheet 4

Robin Schmidt
Probabilistic Inference & Learning

November 11, 2018

## Parametric Regression

### 1. Least Squares Estimation a)

*Proof.* The maximum likelihood estimator $w$ is given by the ordinary least-sqaures estimate:

To show this, we start off by using the formular for the multivariate Gaussian and plugging in the numbers for $p(y|X,w) = \mathcal{N}(y; \phi_x^T w, \sigma^2 I_n)$ and calculating the $log\ p(y|X,w)$:

$$\mathcal{N}(x|\mu,\Sigma) = \frac{1}{2\pi^{\frac{p}{2}}\sqrt{det(\Sigma)}} \cdot exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$$

$$p(y|X,w) = \prod_{i=1}^{N} \frac{1}{2\pi^{\frac{p}{2}}\sqrt{det(\sigma^2 I_n)}} \cdot exp(-\frac{1}{2}(y_i-\phi_x^T w)^T(\sigma^2 I_n)^{-1}(y_i-\phi_x^T w))$$

$$log\ p(y|X,w) = \sum_{i=1}^{N} log(1) - log(2\pi^{\frac{p}{2}}) + log(\sqrt{det(\sigma^2 I_n)}) - \frac{1}{2}(y_i-\phi_x^T w)^T(\sigma^2 I_n)^{-1}(y_i-\phi_x^T w)$$

$$log\ p(y|X,w) = \sum_{i=1}^{N} -\frac{p}{2}log(2\pi) + \frac{1}{2}log(det(\sigma^2 I_n)) - \frac{1}{2}(y_i-\phi_x^T w)^T(\sigma^2 I_n)^{-1}(y_i-\phi_x^T w)$$

$$log\ p(y|X,w) = -\frac{p}{2}log(2\pi) + \frac{1}{2}log(det(\sigma^2 I_n)) - \sum_{i=1}^{N}\frac{(y_i-\phi_x^T w)^2}{2(\sigma^2 I_n)^{-1}}$$

$$log\ p(y|X,w) = -\frac{p}{2}log(2\pi) + \frac{1}{2}log(det(\sigma^2 I_n)) - \frac{1}{2(\sigma^2 I_n)^{-1}}\sum_{i=1}^{N}(y_i-\phi_x^T w)^2$$

If we remove all terms that do not contain the parameter $w$ since these terms do not affect the solution of the optimization problem for the maximum likelihood we get the following expressions, hence the maximum likelihood estimate is defined as the value which makes observed outputs as likely as possible:

$$\begin{aligned} w_{ML} = \underset{w}{arg\ max}\ (log\ p(y|X,w)) &= \underset{w}{arg\ min}\ \sum_{i=1}^{N}(y_i - \sum_{i=1}^{N} w[\phi_x]_i)^2 \\ &= \underset{w}{arg\ min}\ \sum_{i=1}^{N}(y_i - \phi_x^T w)^2 \end{aligned}$$

which is exactly the least squares problem and therefore shows that the maximum likelihood estimator is given by the ordinary least-squares estimate.

Taking the derivative and setting it to zero gives:

$$\begin{gathered} -2\phi_x^T(y - \phi_x^T w) = 0 \\ w_{ML} = (\phi_x \phi_x^T)^{-1}\phi_x y \end{gathered}$$

□

## 1. Least Squares Estimation b)

*Proof.* The maximum a-posteriori estimator is identical to the posterior mean:

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

Because of the fact that $p(y|X)$ does not depend on $w$ we can drop it and get:

$$\begin{aligned} w_{MAP} = arg\,\underset{w}{max}\; p(w|y) &= arg\,\underset{w}{max}\; p(y|X,w)p(w) \\ &= arg\,\underset{w}{max}\; (log\; p(y|X,w) + log\; p(w))) \end{aligned}$$

We can see that the only difference to $w_{ML}$ is in the addition of the logarithm of the prior propability $p(w)$. Using now a particular choice of $\mu = 0$ and $\Sigma = I_F$ for the prior probability and the solution from above leads us to:

$$\begin{aligned} w_{MAP} &= arg\,\underset{w}{max}\; \sum_{i=1}^{N} log(\mathcal{N}(y; \phi_x^T w, \sigma^2 I_n)) + \sum_{i=1}^{F} log(\mathcal{N}(w; 0, I_F)) \\ &= arg\,\underset{w}{max}\; -\frac{1}{2\sigma^2 I_n}\sum_{i=1}^{N}(y_i - \phi_x^T w)^2 - \sum_{i=1}^{F}\frac{1}{2}(w-\mu)^T\Sigma^{-1}(w-\mu) \\ &= arg\,\underset{w}{min}\; \frac{1}{2\sigma^2 I_n}\sum_{i=1}^{N}(y_i - \phi_x^T w)^2 + \sum_{i=1}^{F}\frac{1}{2}(w-\mu)^T\Sigma^{-1}(w-\mu) \\ &= arg\,\underset{w}{min}\; \frac{1}{2\sigma^2 I_n}\sum_{i=1}^{N}(y_i - \phi_x^T w)^2 + \sum_{i=1}^{F}\frac{1}{2}(w-\mu)^T\Sigma^{-1}(w-\mu) \\ &= arg\,\underset{w}{min}\; \sum_{i=1}^{N}(y_i - \phi_x^T w)^2 + \sigma^2 I_n I_F \sum_{i=1}^{F} w_i^2 \\ &= arg\,\underset{w}{min}\; \sum_{i=1}^{N}(y_i - \phi_x^T w)^2 + \sigma^2 I_n I_F ||w||_2^2 \end{aligned}$$

With $log(\mathcal{N}(w;\mu,\Sigma))$ being computed the following way and dropping the terms not containing $w$:

$$\begin{aligned} \mathcal{N}(w;\mu,\Sigma) &= \frac{1}{2\pi^{\frac{p}{2}}\sqrt{det(\Sigma)}} \cdot exp(-\frac{1}{2}(w-\mu)^T\Sigma^{-1}(w-\mu)) \\ log(\mathcal{N}(w;\mu,\Sigma)) &= log(\frac{1}{2\pi^{\frac{p}{2}}\sqrt{det(\Sigma)}}) \cdot -(\frac{1}{2}(w-\mu)^T\Sigma^{-1}(w-\mu)) \end{aligned}$$

Taking the derivative and setting it zero gives:

$$\begin{aligned} -2\phi_x^T(y - \phi_x^T w) + 2\sigma^2 I_n I_F w = 0 \\ w_{MAP} = (\sigma^2 I_n I_F + \phi_x \phi_x^T)^{-1}\phi_x y \end{aligned}$$

This is also shown by just recalling that the posterior distribution $p(w|y)$ is Gaussian. Since the maximum value of a Gaussian probability density function is at its mean, it needs to be the solution to the maximum a-posteriori problem. Therefore the maximum a-posteriori estimator needs to be identical to the posterior mean. □

## 2. Type-II maximum likelihood a)

$$log\ p(y|\phi_x) = log\ (\frac{1}{2\pi^{\frac{n}{2}}\sqrt{det(\phi_x^T\Sigma\phi_x + \sigma^2 I)}} \cdot exp(-\frac{1}{2}(y - \phi_x^T\mu)^T(\phi_x^T\Sigma\phi_x + \sigma^2 I)^{-1}(y - \phi_x^T\mu)))$$

Using the variables $M = \phi_x^T\mu$ and $G = \phi_x^T\Sigma\phi_x + \sigma^2 I$ leads to:

$$log\ p(y|\phi_x) = log\ (\frac{1}{2\pi^{\frac{n}{2}}\sqrt{det(G)}} \cdot exp(-\frac{1}{2}(y - M)^T(G)^{-1}(y - M)))$$

See code for implementation.