TU München
Fakultät für Informatik
PD Dr. Rudolph Triebel
John Chiotellis, Maximilian Denninger

Machine Learning for Computer Vision
Winter term 2018

8. November 2018
Topic: Regression

Exercise 1: Basic model

Consider a linear regression model with basis functions $\phi(x)$ as presented in the lecture. Suppose we have observed N data points $\{x_i, t_i\}_{i=1\ldots N}$.

a) What do we need to estimate if we want to fit this model?

The weights of the basis functions, therefore a vector $\boldsymbol{w} \in \mathbb{R}^m$*, assuming* $\phi(x) \in \mathbb{R}^m$*.*

b) What would be the optimal solution in the sense of sum-of-squares error?

Minimizing the sum-of-squares error amounts to:

$$\boldsymbol{w}_{SSE} = \arg\min_{\boldsymbol{w}} \underbrace{\frac{1}{N}\sum_{i=1}^{N}(t_i - w^T\phi(x_i))^2}_{L(\boldsymbol{w})} \tag{1}$$

We take the gradient of the loss function w.r.t. $\boldsymbol{w}$*, set it to 0 and solve for* $\boldsymbol{w}$:

$$\begin{aligned}
\nabla_{\boldsymbol{w}} L(\boldsymbol{w}) &= \nabla_{\boldsymbol{w}} \left(\frac{1}{N}\sum_{i=1}^{N}(t_i - w^T\phi(x_i))^2\right) \\
&= \nabla_{\boldsymbol{w}} \left(\frac{1}{2}\sum_{i=1}^{N}(t_i - w^T\phi(x_i))^2\right) \\
&= \nabla_{\boldsymbol{w}} \left(\frac{1}{2}\sum_{i=1}^{N} -2t_i w^T\phi(x_i) + ||w^T\phi(x_i)||^2\right) \\
&= \sum_{i=1}^{N} -t_i\phi(x_i) + (w^T\phi(x_i))\phi(x_i) \overset{!}{=} 0
\end{aligned}$$

Solving for $\boldsymbol{w}$:

$$\begin{aligned}
\sum_{i=1}^{N}(w^T\phi(x_i))\phi(x_i) &= \sum_{i=1}^{N} t_i\phi(x_i)\\
\sum_{i=1}^{N}\phi(x_i)(\phi(x_i)^T w) &= \sum_{i=1}^{N} t_i\phi(x_i)\\
(\sum_{i=1}^{N}\phi(x_i)(\phi(x_i)^T)w &= \sum_{i=1}^{N} t_i\phi(x_i)\\
\Phi^T\Phi w &= \Phi^T t\\
\boldsymbol{w}_{SSE} &= (\Phi^T\Phi)^{-1}\Phi^T t
\end{aligned}$$

c) Can you define and solve the problem in a probabilistic way?
Hint: You have to make some assumptions.

To define the problem in a probabilistic way, we have to assume a distribution over our observed target variables. The simplest way is to choose this distribution to be Gaussian, centered at our model's predictions with some deviation σ. So we model our observations as the likelihood:

$$p(t_i|x_i, w) = \mathcal{N}(t_i|w^T\phi(x_i), \sigma^2)$$

We also assume that the data points we observed are independent and identically distributed (i.i.d.). Therefore the likelihood over all points is just the product of the likelihoods of each observation:

$$p(\boldsymbol{t}|\boldsymbol{X}, \boldsymbol{w}) = \prod_{i=1}^{N} p(t_i|x_i, w) = \prod_{i=1}^{N} \mathcal{N}(t_i|w^T\phi(x_i), \sigma^2)$$

We want to find the weights that maximize this likelihood:

$$\boldsymbol{w}_{ML} = \arg\max_{\boldsymbol{w}} p(\boldsymbol{t}|\boldsymbol{X}, \boldsymbol{w}) = \arg\max_{\boldsymbol{w}} \prod_{i=1}^{N} \mathcal{N}(t_i|w^T\phi(x_i), \sigma^2)$$

Now we make use of two facts:

a) Maximizing a function $f(x)$ is the same as maximizing $\log f(x)$ because the logarithm is a monotonically increasing function.

b) Maximizing a function $f(x)$ is the same as minimizing $-f(x)$.

Using these two, we end up minimizing the negative log-likelihood (NLL):

$$\begin{aligned}
\boldsymbol{w}_{ML} = \boldsymbol{w}_{NLL} &= \arg\min_{\boldsymbol{w}} - \ln\left(\prod_{i=1}^{N} \mathcal{N}(t_i|w^T\phi(x_i), \sigma^2)\right) \\
&= \arg\min_{\boldsymbol{w}} - \sum_{i=1}^{N} \ln \mathcal{N}(t_i|w^T\phi(x_i), \sigma^2) \\
&= \arg\min_{\boldsymbol{w}} - \sum_{i=1}^{N} \ln\left((2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(t_i - w^T\phi(x_i))^2}{2\sigma^2}\right)\right) \\
&= \arg\min_{\boldsymbol{w}} - \sum_{i=1}^{N} \ln(2\pi\sigma^2)^{-1/2} + \ln\left(\exp\left(-\frac{(t_i - w^T\phi(x_i))^2}{2\sigma^2}\right)\right) \\
&= \arg\min_{\boldsymbol{w}} -N\ln(2\pi\sigma^2)^{-1/2} - \sum_{i=1}^{N}\left(-\frac{(t_i - w^T\phi(x_i))^2}{2\sigma^2}\right) \\
&= \arg\min_{\boldsymbol{w}} \frac{1}{2\sigma^2}\sum_{i=1}^{N}(t_i - w^T\phi(x_i))^2 = \arg\min_{\boldsymbol{w}} \frac{1}{2}\sum_{i=1}^{N}(t_i - w^T\phi(x_i))^2 = \boldsymbol{w}_{SSE}
\end{aligned}$$

Therefore, we ended up with the same solution: $\boldsymbol{w}_{ML} = \boldsymbol{w}_{NLL} = \boldsymbol{w}_{SSE} = (\Phi^T\Phi)^{-1}\Phi^T t$.

Exercise 2: Bayesian Update

Now we assume a Gaussian prior distribution for the weights:

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|m_0, S_0)$$

Suppose we have already observed N data points, so the posterior distribution is

$$p(\boldsymbol{w}|\mathbf{t}) = \mathcal{N}(\boldsymbol{w}|m_N, S_N)$$

with

$$m_N = S_N(S_0^{-1}m_0 + \sigma^{-2}\Phi^T\mathbf{t}) \quad \text{and} \quad S_N^{-1} = S_0^{-1} + \sigma^{-2}\Phi^T\Phi.$$

Now, we observe a new data point (x_{N+1}, t_{N+1}). What is the new posterior?

Using Bayes rule, we found out that having a Gaussian prior and a Gaussian likelihood gave us a Gaussian posterior which we can use as the prior for the next iteration (next sample that we observe). Now we want to compute $p(\boldsymbol{w}|\mathbf{t}, t_{N+1}, x_{N+1})$ which reduces to $p(\boldsymbol{w}|t_{N+1}, x_{N+1}, m_N, S_N)$.

Our likelihood is

$$p(t_{N+1}|x_{N+1}, \boldsymbol{w}) = \mathcal{N}(t_{N+1}|y(\boldsymbol{w}, \phi(x_{N+1})), \sigma^2)$$

Let $\phi_N = \phi(x_N)$ to simplify notation. Writing the likelihood explicitly we get

$$p(t_{N+1}|x_{N+1}, \boldsymbol{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t_{N+1} - \boldsymbol{w}^T\phi_{N+1})^2}{2\sigma^2}\right)$$

Our posterior is

$$p(\boldsymbol{w}|t_{N+1}, x_{N+1}, m_N, S_N) = \frac{p(t_{N+1}|x_{N+1}, \boldsymbol{w})p(\boldsymbol{w}|\mathbf{t})}{p(t_{N+1}|x_{N+1}, \mathbf{t})}$$

We want the maximum likelihood of the posterior, namely the maximum a posteriori (MAP). The denominator is independent of $\boldsymbol{w}$ so for now we can ignore it.

$$\begin{aligned} p(\boldsymbol{w}|t_{N+1}, x_{N+1}, m_N, S_N) &\propto p(\boldsymbol{w}|\mathbf{t})p(t_{N+1}|x_{N+1}, \boldsymbol{w}) \\ &\propto \exp\left(-\frac{1}{2}(\boldsymbol{w} - m_N)^T S_N^{-1}(\boldsymbol{w} - m_N) - \frac{(t_{N+1} - \boldsymbol{w}^T\phi_{N+1})^2}{2\sigma^2}\right) \end{aligned}$$

Therefore we have

$$\begin{aligned} \boldsymbol{w}_{MAP} &= \arg\max_w p(\boldsymbol{w}|t)p(t_{N+1}|x_{N+1}, \boldsymbol{w}) = \arg\min_w - \ln\left(p(\boldsymbol{w}|t)p(t_{N+1}|x_{N+1}, \boldsymbol{w})\right) \\ &= \arg\min_w (\boldsymbol{w} - m_N)^T S_N^{-1}(\boldsymbol{w} - m_N) + \frac{(t_{N+1} - \boldsymbol{w}^T\phi_{N+1})^2}{\sigma^2} \\ &= \arg\min_w \boldsymbol{w}^T S_N^{-1}\boldsymbol{w} - 2\boldsymbol{w}^T S_N^{-1} m_N - 2\frac{\boldsymbol{w}^T\phi_{N+1}t_{N+1}}{\sigma^2} + \frac{\boldsymbol{w}^T\phi_{N+1}\phi_{N+1}^T\boldsymbol{w}}{\sigma^2} \\ &= \arg\min_w \boldsymbol{w}^T(S_N^{-1} + \frac{\phi_{N+1}\phi_{N+1}^T}{\sigma^2})\boldsymbol{w} - 2\boldsymbol{w}^T\left(S_N^{-1}m_N + \frac{\phi_{N+1}t_{N+1}}{\sigma^2}\right) \end{aligned}$$

where we have ignored all terms that are independent of $\boldsymbol{w}$.
Now we can save us some time and computation by observing how the maximum likelihood of our prior relates to the maximum likelihood of our posterior. For our prior we had:

$$\begin{aligned} \boldsymbol{w}_{ML} &= \arg\max_w p(\boldsymbol{w}|t) = \arg\min_w - \ln p(\boldsymbol{w}|t) = \arg\min_w - \ln \mathcal{N}(\boldsymbol{w}|m_N, S_N) \\ &= \arg\min_w - \ln(2\pi|S_N|)^{-\frac{d}{2}} + \frac{1}{2}(\boldsymbol{w} - m_N)^T S_N^{-1}(\boldsymbol{w} - m_N) \\ &= \arg\min_w (\boldsymbol{w} - m_N)^T S_N^{-1}(\boldsymbol{w} - m_N) = \arg\min_w \boldsymbol{w}^T S_N^{-1}\boldsymbol{w} - 2\boldsymbol{w}^T S_N^{-1} m_N \end{aligned}$$

We observe that for our prior, the inverse covariance matrix for N points S_N^{-1} stands alone in the second-order term of the weights. This means, what stands alone in the second-order term of the weights of the posterior, should be the inverse covariance matrix for $N + 1$ points, S_{N+1}^{-1}:

$$S_{N+1}^{-1} = S_N^{-1} + \frac{1}{\sigma^2}\phi_{N+1}\phi_{N+1}^T$$

This means we can update the (inverse) covariance matrix with a rank-one update based on the one new point we observed. Equivalently, for the first-order term we have:

$$S_{N+1}^{-1}m_{N+1} = S_N^{-1}m_N + \frac{1}{\sigma^2}\phi_{N+1}t_{N+1} \Rightarrow m_{N+1} = S_{N+1}(S_N^{-1}m_N + \frac{\phi_{N+1}t_{N+1}}{\sigma^2})$$