Bayesian Machine Learning
Course 67564
Solution To Exercise 1: Bayesian Statistics and Gaussians

Barak Haim 0

10/11/2022

Contents

1 Bayesian Statistics

1.1 MSE and BMSE

1.1.1 Q1 MSE $\left[\hat{\theta}\right]=bais^{2}\left(\hat{\theta}\right)+var\left[\hat{\theta}\right]$

Note

$$bais^{2}\left(\hat{\theta}\right)=\left(E_{D}\left[\hat{\theta}(D)\right]-\theta\right)^{2}=E_{D}\left[\hat{\theta}(D)\right]^{2}-2E_{D}\left[\hat{\theta}(D)\right]\theta+\theta^{2}$$

and so:

$$bais^{2}\left(\hat{\theta}\right)+var\left[\hat{\theta}\right]=E_{D}\left[\hat{\theta}(D)^{2}\right]-2E_{D}\left[\hat{\theta}(D)\right]\theta+\theta^{2}$$

Moreover:

$$\begin{aligned}\left\Vert \theta-\hat{\theta}(D)\right\Vert ^{2}&=\left\langle \theta-\hat{\theta}(D),\theta-\hat{\theta}(D)\right\rangle \\ &=\left\langle \theta,\theta-\hat{\theta}(D)\right\rangle -\left\langle \hat{\theta}(D),\theta-\hat{\theta}(D)\right\rangle \\ &=\left\langle \theta,\theta\right\rangle -\left\langle \theta,\hat{\theta}(D)\right\rangle -\left\langle \hat{\theta}(D),\theta\right\rangle +\left\langle \hat{\theta}(D),\hat{\theta}(D)\right\rangle \\ &=\left\langle \theta,\theta\right\rangle -2\left\langle \theta,\hat{\theta}(D)\right\rangle +\left\langle \hat{\theta}(D),\hat{\theta}(D)\right\rangle \\ &=\left\langle \hat{\theta}(D),\hat{\theta}(D)\right\rangle -2\left\langle \theta,\hat{\theta}(D)\right\rangle +\left\langle \theta,\theta\right\rangle \end{aligned}$$

Now, from linearity of E:

$$E_{D}\left\Vert \theta-\hat{\theta}(D)\right\Vert ^{2}=E_{D}\left[\hat{\theta}(D)^{2}\right]-2E_{D}\left[\left\langle \theta,\hat{\theta}(D)\right\rangle \right]+E_{D}\left[\theta^{2}\right]$$

Because θ is constant with regards to D, E_{D} is not effected by θ and so

$$E_{D}\left[\left\Vert \theta-\hat{\theta}(D)\right\Vert ^{2}\right]=E_{D}\left[\hat{\theta}(D)^{2}\right]-2E_{D}\left[\hat{\theta}(D)\right]\theta+\theta^{2}=bais^{2}\left(\hat{\theta}\right)+var\left[\hat{\theta}\right]$$

Q.E.D.

1.1.2 Q2 $\hat{\theta}_{a}=a\cdot\underset{\tilde{\theta}}{argmin}\sum_{i=1}^{N}\left(y_{i}-\tilde{\theta}\right)^{2}$, $MSE\left[\hat{\theta}_{a}\right]=?$

First we can compute $\hat{\theta}_{a}$by finding the minima of $MD(\tilde{\theta})\triangleq\sum_{i=1}^{N}\left(y_{i}-\tilde{\theta}\right)^{2}$ (we know there is such a unique term as $MD(\tilde{\theta})$ is convex). So:

$$\frac{\partial MD(x)}{\partial x}=\frac{\partial}{\partial x}\sum_{i=1}^{N}\left(y_{i}-x\right)^{2}=\sum_{i=1}^{N}\frac{\partial}{\partial x}\left(y_{i}-x\right)^{2}=-2\sum_{i=1}^{N}\left(y_{i}-x\right)=2\sum_{i=1}^{N}\left(x-y_{i}\right)=2\left(Nx-\sum_{i=1}^{N}y_{i}\right)$$

So, $\frac{\partial MD(x)}{\partial x}=0$ iff $Nx-\sum_{i=1}^{N}y_{i}=0$ iff $x=\sum_{i=1}^{N}\frac{y_{i}}{N}$. We get:

$$\hat{\theta}_{a}=a\cdot\underset{\tilde{\theta}}{argmin}\sum_{i=1}^{N}\left(y_{i}-\tilde{\theta}\right)^{2}=\frac{a}{N}\sum_{i=1}^{N}y_{i}$$

Now:

$$E_D\left[\hat{\theta}_a(D)\right] = E_D\left[\frac{a}{N}\sum_{i=1}^{N} y_i\right] = \frac{a}{N}\sum_{i=1}^{N} E_D\left[y_i\right]$$

Because $y_i's$ are sampled from θ we get $E_D\left[y_i\right] = \theta$ for each i and so:

$$E_D\left[\hat{\theta}_a(D)\right] = \frac{a}{N}\sum_{i=1}^{N} \theta = \frac{a}{N} N\theta = a\theta$$

And the bias is:

$$bais\left(\hat{\theta}\right) = E_D\left[\hat{\theta}_a(D)\right] - \theta = a\theta - \theta = \theta\left(a-1\right)$$

And so:

$$bais^2\left(\hat{\theta}\right) = \theta^2\left(a-1\right)^2$$

And the variance is:

$$E_D\left[\left(\hat{\theta}_a(D)\right)^2\right] = E_D\left[\left(\frac{a}{N}\sum_{i=1}^{N} y_i\right)^2\right] = \left(\frac{a}{N}\right)^2 E_D\left[\left(\sum_{i=1}^{N} y_i\right)^2\right]$$

$$var\left[\hat{\theta}_a(D)\right] = var\left[\frac{a}{N}\sum_{i=1}^{N} y_i\right] = \left(\frac{a}{N}\right)^2 var\left[\sum_{i=1}^{N} y_i\right] = \left(\frac{a}{N}\right)^2 \sum_{i=1}^{N} var\left[y_i\right]$$

Where the last transition is due to the fact the y_i's ar disjoint. Since $y_i \sim N\left(\theta, \sigma^2\right)$ we get $var\left[y_i\right] = \sigma_2$ For each i. Hence:

$$var\left[\hat{\theta}_a(D)\right] = \frac{a^2}{N^2} N\sigma^2 = \frac{a^2\sigma^2}{N}$$

Now, using Q1:

$$\begin{aligned}\mathbf{MSE}\left[\hat{\theta}\right] &= bais^2\left(\hat{\theta}\right) + var\left[\hat{\theta}\right]\\ &= \theta^2\left(a-1\right)^2 + \frac{a^2\sigma^2}{N}\end{aligned}$$

As MSE $\left[\hat{\theta}\right]$ depends on a we're left to determine which a is optimal, if any. We look for the extrama:

$$\begin{aligned}\frac{\partial}{\partial a}\mathbf{MSE}\left[\hat{\theta}\right] &= \frac{\partial}{\partial a}\theta^2\left(a-1\right)^2 + \frac{\partial}{\partial a}\frac{a^2\sigma^2}{N} = \theta^2\frac{\partial}{\partial a}\left(a-1\right)^2 + \frac{\sigma^2}{N}\frac{\partial}{\partial a}a^2\\ &= \theta^2\cdot 2\left(a-1\right) + \frac{\sigma^2}{N}\cdot 2a = 2a\theta^2 - 2\theta^2 + \frac{2a\sigma^2}{N} = 2a\left(\theta^2 + \frac{\sigma^2}{N}\right) - 2\theta^2\end{aligned}$$

Now $\frac{\partial}{\partial a}$MSE $\left[\hat{\theta}\right] = 0$ iff

$$a^{MSE} = \frac{2\theta^2}{2\left(\theta^2 + \frac{\sigma^2}{N}\right)} = \frac{\theta^2}{N\theta^2 + \sigma^2}\cdot\frac{1}{\frac{1}{N}} = \frac{N\theta^2}{N\theta^2 + \sigma^2}$$

Since a^{MSE} is a function of θ, we can't say a single value of a is globaly optimal.

1.1.3 Q3 BMSE $\left[\hat{\theta}\right] = \int p(\theta), \textbf{MSE}\left[\hat{\theta}\right] d\theta$, $\theta \sim N(0,1)$

Note

$$\textbf{BMSE}\left[\hat{\theta}\right] = E_\theta\left[\textbf{MSE}\left[\hat{\theta}\right]\right] = E_\theta\left[\theta^2 (a-1)^2 + \frac{a^2\sigma^2}{N}\right] = (a-1)^2 E_\theta\left[\theta^2\right] + \frac{a^2\sigma^2}{N}$$

Also, as we can write $p(\theta)$ explicitly and use Wolfram Alpha, we get - $E_\theta[\theta^2] = \int p(\theta)\theta^2 d\theta = 1$, than:

$$\textbf{BMSE}\left[\hat{\theta}\right] = (a-1)^2 + \frac{a^2\sigma^2}{N}$$

Again, we look for the max by a:

$$\begin{aligned}\frac{\partial}{\partial a}\textbf{BMSE}\left[\hat{\theta}\right] &= \frac{\partial}{\partial a}(a-1)^2 + \frac{\partial}{\partial a}\frac{a^2\sigma^2}{N} \\ &= 2a - 2 + \frac{2a\sigma^2}{N} = 2a\left(1 - \frac{\sigma^2}{N}\right) - 2 \stackrel{?}{=} 0 \\ &\iff a^{MMSE} = \frac{1}{\left(1 - \frac{\sigma^2}{N}\right)} = \frac{N}{(N-\sigma^2)}\end{aligned}$$

Hence we can find an optimal a regardless of θ, i.e. a globally optimal a.

1.2 Prior, Likelihood and Posterior

1.2.1 Q4 $\theta \sim U[a,b]$, $y|\theta \sim U([\theta-\delta, \theta+\delta])$

So for a single data point y, we get $p(y|\theta) = \frac{1}{2\delta}$ and - $p(\theta) = \frac{1}{b-a}$. Using Bayes' law:

$$\begin{aligned}p(\theta|y) &= \frac{1}{c}\cdot\begin{cases}\frac{p(y|\theta)}{b-a} & \theta \in [a,b] \\ 0 & else\end{cases} = \frac{1}{c}\cdot\begin{cases}\frac{1}{2\delta(b-a)} & \theta \in [a,b]\, and\, y \in [\theta-\delta, \theta+\delta] \\ 0 & else\end{cases} \\ &= \frac{1}{c}\cdot\begin{cases}\frac{1}{2\delta(b-a)} & a \le \theta \le b\, and\, \theta-\delta \le y \le \theta+\delta \\ 0 & else\end{cases} = \bigstar\end{aligned}$$

As $\theta - \delta \le y \le \theta + \delta$ iff $y - \delta \le \theta \le y + \delta$ we can rewrite the condition for $p(\theta|y) \neq 0$ as: $min\{a, y-\delta\} \le \theta \le max\{y+\delta, b\}$. Hence:

$$\bigstar = \frac{1}{c}\cdot\begin{cases}\frac{1}{2\delta(b-a)} & min\{a, y-\delta\} \le \theta \le max\{y+\delta, b\} \\ 0 & else\end{cases}$$

Now we know the PDF function holds the condition: $\int\limits_{-\infty}^{\infty} p(\theta|y)d\theta = 1$ so c above must be a normliztion factor in the range $[min\{a, y-\delta\}, max\{y+\delta, b\}]$ so we can say $p(\theta|y)$ is continues uniform in the range $[min\{a, y-\delta\}, max\{y+\delta, b\}]$, i.e.:

$$p(y|\theta) \sim U(\theta|[min\{a, y-\delta\}, max\{y+\delta, b\}])$$

1.2.2 Q5 $\theta \sim U[a,b],\ y|\theta \sim N\left(\theta, \lambda^2\right)$

We have - $p(\theta) = \frac{1}{b-a}$ **and** $p(y|\theta) = \frac{1}{\sqrt{2\pi\lambda^2}} \exp\left(-\frac{1}{2\lambda^2}\left\|y-\theta\right\|^2\right)$**. Using Bayes' law:**

$$p(\theta|y) = \frac{1}{c} \cdot \begin{cases} \frac{1}{b-a} \cdot \frac{1}{\sqrt{2\pi\lambda^2}} \exp\left(-\frac{1}{2\lambda^2}\left\|y-\theta\right\|^2\right) & \theta \in [a,b] \\ 0 & else \end{cases}$$

where $c = P(y) = \int_\theta p(\theta)p(y|\theta)d\theta$ **and because** $p(\theta) = 0$ **outside the range** $[a,b]$ **we get:**

$$c = \frac{1}{b-a}\int_a^b p(y|\theta)d\theta = \frac{1}{b-a}\int_a^b \frac{1}{\sqrt{2\pi\lambda^2}} \exp\left(-\frac{1}{2\lambda^2}\left\|y-\theta\right\|^2\right)d\theta$$

Thus we can write

$$p(\theta|y) = \begin{cases} \frac{N\left(\theta|y,\lambda^2\right)}{\int_a^b \frac{1}{\sqrt{2\pi\lambda^2}} \exp\left(-\frac{1}{2\lambda^2}\left\|y-\theta\right\|^2\right)d\theta} & \theta \in [a,b] \\ 0 & else \end{cases}$$

1.2.3 Q6 $\theta \sim N\left(\mu, \sigma^2\right),\ y|\theta \sim N\left(h\cdot\theta, \lambda^2\right)$

We have - $p(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left\|\theta-\mu\right\|^2\right)$ **and** $p(y|\theta) = \frac{1}{\sqrt{2\pi\lambda^2}} \exp\left(-\frac{1}{2\lambda^2}\left\|y-h\theta\right\|^2\right)$**. Using Bayes' law:**

$$\begin{aligned} p(\theta|y) &= \frac{1}{c}\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left\|\theta-\mu\right\|^2\right)\frac{1}{\sqrt{2\pi\lambda^2}} \exp\left(-\frac{1}{2\lambda^2}\left\|y-h\theta\right\|^2\right) \\ &= \frac{1}{c}\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \frac{1}{\sqrt{2\pi\lambda^2}} \exp\left(-\frac{1}{2\sigma^2}\left\|\theta-\mu\right\|^2 - \frac{1}{2\lambda^2}\left\|y-h\theta\right\|^2\right) \\ &= \frac{1}{z}\exp\left(-\Delta\right) \end{aligned}$$

As Δ**is a quadratic term of** θ **and we saw in the recitation this idicates** $p(\theta|y)$ **is indeed a Gaussian, we're left to find its mean and variance and can use the derivitive trick to do so. So:**

$$\begin{aligned} \frac{\partial\Delta}{\partial\theta} &= \frac{1}{\sigma^2}\left(\theta-\mu\right) - \frac{h}{\lambda^2}\left(y-h\theta\right) \\ &= \frac{\lambda^2\theta - \lambda^2\mu - \sigma^2hy + \sigma^2h^2\theta}{\left(\sigma\lambda\right)^2} \\ &= \frac{\left(\lambda^2+\sigma^2h^2\right)\theta - \lambda^2\mu - \sigma^2hy}{\left(\sigma\lambda\right)^2} \\ &= \frac{\left(\lambda^2+\sigma^2h^2\right)}{\left(\sigma\lambda\right)^2}\left(\theta - \frac{\lambda^2\mu + \sigma^2hy}{\left(\lambda^2+\sigma^2h^2\right)}\right) \end{aligned}$$

Hence -

$$p(\theta|y) \sim N\left(\frac{\left(\lambda^2\mu + \sigma^2hy\right)}{\left(\lambda^2+\sigma^2h^2\right)}, \left(\frac{\left(\lambda^2+\sigma^2h^2\right)}{\left(\sigma\lambda\right)^2}\right)^{-1}\right)$$

2 Gaussians

2.1 Sampling from a Multivariate Normal

2.1.1 Q7

First note that - $f^{-1}(y) = A^{-1}(y-b)$**. Now:**

$$\frac{\partial}{\partial y}f^{-1}(y) = \frac{\partial}{\partial y}A^{-1}(y-b) = \frac{\partial}{\partial y}(y-b)\cdot\frac{\partial}{\partial(y-b)}A^{-1}(y-b) = 1\cdot A^{-1} = A^{-1}$$

Also:

$$\begin{aligned} p_x(f^{-1}(y)) &= \frac{1}{\sqrt{(2\pi)^d\,|\Sigma|}}\exp\left(-\frac{1}{2}(f^{-1}(y)-\mu)^T\Sigma^{-1}(f^{-1}(y)-\mu)\right) \\ &= \frac{1}{\sqrt{(2\pi)^d\,|\Sigma|}}\exp\left(-\frac{1}{2}(A^{-1}(y-b)-\mu)^T\Sigma^{-1}(A^{-1}(y-b)-\mu)\right) \end{aligned}$$

So togther:

$$\begin{aligned} p_y(y) &= p_x(f^{-1}(y))\cdot\left|\frac{\partial}{\partial y}f^{-1}(y)\right| \\ &= \frac{|A|}{\sqrt{(2\pi)^d\,|\Sigma|}}\exp\left(-\frac{1}{2}(A^{-1}(y-b)-\mu)^T\Sigma^{-1}(A^{-1}(y-b)-\mu)\right) \end{aligned}$$

Denote $\Delta = -\frac{1}{2}(A^{-1}(y-b)-\mu)^T\Sigma^{-1}(A^{-1}(y-b)-\mu)$**. As we saw in recitation, in order to show** $y \propto N(\mu_y, \Sigma_y)$ **it's enough to show ... From the chain role:**

$$\frac{\partial}{\partial y}\Delta = \frac{\partial f^{-1}(y)}{\partial y}\cdot\frac{\partial}{\partial f^{-1}(y)}\Delta$$

We know Δ **is the term for the Mahalanobis distance and so** $\frac{\partial}{\partial f^{-1}(y)}\Delta = \Sigma^{-1}(f^{-1}(y)-\mu)$ **and we already saw** $\frac{\partial}{\partial y}f^{-1}(y) = \left(A^{-1}\right)^T = \left(A^T\right)^{-1}$**. Togther we get:**

$$\begin{aligned} \frac{\partial}{\partial y}\Delta &= \left(A^{-1}\right)^T\Sigma^{-1}(f^{-1}(y)-\mu) = \left(A^{-1}\right)^T\Sigma^{-1}(A^{-1}(y-b)-\mu) \\ &= \left(A^{-1}\right)^T\Sigma^{-1}A^{-1}(y-b-A\mu) \end{aligned}$$

So $\mu_y = b - A\mu$ **and** $\Sigma_y = A\Sigma A^T$**. Lastly, we want to make sure** Σ_y **is indeed a PD. As** A **is invertible, it is infact a homomorphism from** $\mathbb{R}^n$ **onto itself. Specificly, A has a decomposition to eignvalues** $\{a_i\}_{i=1}^n$ **and the eigenvalues of** Σ_y **are** $\{a_i^2\lambda_i\}_{i=1}^n$**. Since for each i** $0 < a_i^2, \lambda_i$ **we get that** Σ_y **is indeed a PD matrix. Hence we can write:**

$$\frac{\partial}{\partial y}\Delta = \Sigma_y(y-\mu_y)$$

And so we get that $y \sim N(\mu_y, \Sigma_y)$ **is a Gaussian.**

2.1.2 Q8

In the terms of Q7 - we have x as y $\sim N(\mu_x, \Sigma_x)$, A is R for $\Sigma_x = RR^T$, b as μ. we get $\Sigma_x = R\Sigma_z R^T = RIR^T = \Sigma$ and $\mu_x = b - A\mu_z = \mu - 0$. If we pretend the double meaning of x,y,z here were not confusing, we get that by using $f(z) = Rz + \mu$ we're able to move from $N(0, I)$ To $N(\mu, \Sigma)$■

2.2 Product of Gaussians

$$\begin{aligned} x &\sim N(\mu, \Sigma) \\ \eta &\sim N(0, \Gamma) \\ y &= Hx + \eta \end{aligned}$$

2.2.1 Q9 $p(y|x)$

For a fixed x, y is an affine transformation of η. Hence - $p(y|x) \sim N(Hx, \Gamma)$.

2.2.2 Q10 $p(y)$

$$\begin{aligned} p(x)p(y|x) &= \frac{1}{z_1} \exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right] \cdot \frac{1}{z_2} \exp\left[-\frac{1}{2}(y-Hx)^T\Gamma^{-1}(y-Hx)\right] \\ &= \frac{1}{z_1 z_2} \exp\left[-\frac{1}{2}\left((x-\mu)^T\Sigma^{-1}(x-\mu) + (y-Hx)^T\Gamma^{-1}(y-Hx)\right)\right] \end{aligned}$$

Now we know $(x-\mu)^T\Sigma^{-1}(x-\mu)^T$ is a quadratic form and hence contains quadratic terms of x. Moreover, $(y-Hx)^T\Gamma^{-1}(y-Hx)^T$ is also a quadratic form, this time of the linear combibation $y - Hx$. Using its linearty we'll get quadratic terms of x,y or their product. As we saw in class, having quadratic terms of the vector (x, y) in the exponent is enough in order to determine $p(x, y) = p(x)p(y|x)$ is also a Guassion. In turn, this implies $p(y)$ is also a Guassion. This infact is enough to describe the distribution as $p(y) \sim N(E[y], var[y])$. Remeber that in exercise 0 we showed $var(y) = H\Sigma H^T + \Gamma$ and from form linearity we get $E[y] = HE[x] + E[\eta] = H\mu + 0$. Thus we get:

$$y \sim N\left(H\mu, H\Sigma H^T + \Gamma\right)$$

2.2.3 Q11 $p(x|y)$

Using Bayes' law we know $p(x|y) \propto p(x)p(y|x)$. As we already know this is a Gaussion, we can use the derivative trick in order to bring the derivative of the term in the exponent the canonical form and deduce the mean and variance. Denote

$$\Delta = \frac{1}{2}\left((x-\mu)^T\Sigma^{-1}(x-\mu) + (y-Hx)^T\Gamma^{-1}(y-Hx)\right)$$

Than:

$$
\begin{aligned}
\frac{\partial \Delta}{\partial x} &= \Sigma^{-1}(x-\mu) - H^T\Gamma^{-1}(y - Hx) \\
&= \Sigma^{-1}x - \Sigma^{-1}\mu - H^T\Gamma^{-1}y + H^T\Gamma^{-1}Hx \\
&= \left(\Sigma^{-1} + H^T\Gamma^{-1}H\right)x - \Sigma^{-1}\mu - H^T\Gamma^{-1}y \\
&= \left(\Sigma^{-1} + H^T\Gamma^{-1}H\right)\left(x - \frac{\Sigma^{-1}\mu + H^T\Gamma^{-1}y}{(\Sigma^{-1} + H^T\Gamma^{-1}H)}\right)
\end{aligned}
$$

Hence:

$$
x \sim N\left[\frac{\Sigma^{-1}\mu + H^T\Gamma^{-1}y}{(\Sigma^{-1} + H^T\Gamma^{-1}H)}, \left(\Sigma^{-1} + H^T\Gamma^{-1}H\right)^{-1}\right]
$$

Q.E.D.