Machine Learning	Prof. Dr. Mario Fritz
Sheet 1 (Solution)	***Tutorial 2022.04.21***

Exercise 1: Fruits

Suppose that we have three coloured boxes r (red), b (blue), and g (green). Box r contains 3 apples, 4 oranges, and 3 limes, box b contains 1 apple, 1 orange, and 0 limes, and box g contains 3 apples, 3 oranges, and 4 limes. A box is chosen at random with probabilities $p(r) = 0.2, p(b) = 0.2, p(g) = 0.6$, and a piece of fruit is selected from the box (with equal probability of selecting any of the items in the selected box).

i) What is the probability of selecting an apple?

ii) If we observe that the selected fruit is in fact an orange, what is the probability that it came from the green box?

Solution:

i) (Application of sum rule and product rule)

$$\begin{aligned}
\mathrm{p(apple)} &= \mathrm{p(apple|r) \cdot p(r) + p(apple|b) \cdot p(b) + p(apple|g) \cdot p(g)} \\
&= \frac{3}{3+4+3} \cdot 0.2 + \frac{1}{1+1+0} \cdot 0.2 + \frac{3}{3+3+4} \cdot 0.6 \\
&= 0.34
\end{aligned}$$

ii) (Definition of conditional probability, Sum rule and product rule)

$$\begin{aligned}
\mathrm{p(g|orange)} &= \frac{\mathrm{p(g, orange)}}{\mathrm{p(orange)}} \\
&= \frac{\mathrm{p(orange|g) \cdot p(g)}}{\mathrm{p(orange|r) \cdot p(r) + p(orange|b) \cdot p(b) + p(orange|g) \cdot p(g)}} \\
&= \frac{0.6 \cdot \frac{3}{3+3+4}}{\frac{4}{3+3+4} \cdot 0.2 + \frac{1}{1+1} \cdot 0.2 + \frac{3}{3+3+4} \cdot 0.6} = 0.5
\end{aligned}$$

Exercise 2: Maximum Density

Consider a probability density $p_x(x)$ defined over a continuous variable x, and suppose that we make a nonlinear change of variable using $x = g(y)$, so that the density transforms according to

$$p_y(y) = p_x(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| = p_x(g(y))|g'(y)| \tag{1}$$

i) By differentiating Equation 1, show that the location $\hat{y}$ of the maximum of the density in y is not in general related to the location $\hat{x}$ of the maximum of the density over x by the simple functional relation $\hat{x} = g(\hat{y})$ as a consequence of the Jacobian factor. This shows that the maximum of a probability density (in contrast to a simple function) is dependent on the choice of variable.

ii) Verify that, in the case of a linear transformation, the location of the maximum transforms in the same way as the variable itself.

Solution:
The main task of this exercise is to investigate whether the following is true:

$$\hat{x} \text{ is a mode of } p_x \text{ and } \hat{x} = g(\hat{y}) \qquad \Longrightarrow \qquad \hat{y} = g^{-1}(\hat{x}) \text{ is a mode of } p_y$$

We first write $g'(y) = s|g'(y)|$ where $s \in \{+1, -1\}$ denotes the sign of $g'(y)$.

Plug in Equation 1, we have: $p_y(y) = p_x(g(y)) \cdot sg'(y)$.

To find the mode (maximum), we need to investigate the derivative (first order condition):

$$p'_y(y) = s \cdot p'_x(g(y)) \cdot (g'(y))^2 + s \cdot p_x(g(y)) \cdot g''(y)$$

As $\hat{x}$ is a mode of p_x, we know that $p'_x(\hat{x}) = 0$ and $p''_x(\hat{x}) < 0$.

Thus, $p'_y(\hat{y}) = s \cdot p_x(g(\hat{y})) \cdot g''(\hat{y})$ as the first term in $p'_y(\hat{y})$ vanishes.

- If $x = g(y)$ is a linear transformation, we have $g''(\hat{y}) = 0$, and thus $p'_y(\hat{y}) = 0$.

  Verifying the second order condition, we have: $p''_y(\hat{y}) = s \cdot p''_x(\hat{x})(g'(\hat{y}))^3 < 0$.

  Thus, $\hat{y}$ is a mode of p_y.

- In general, we do not know the value of $g''(\hat{y})$, and $\hat{y}$ is ***not*** a mode of p_y if $g''(\hat{y}) \neq 0$. This means, the mode of the transformed density p_y depends on the choice of variable. Hence, in general:

  $$\hat{x} \text{ is a mode of } p_x \text{ and } \hat{x} = g(\hat{y}) \qquad \not\Longrightarrow \qquad \hat{y} = g^{-1}(\hat{x}) \text{ is a mode of } p_y$$

  See Figure 1 for an example of a non-linear transformation of the variable. $p_x(x)$ is a Gaussian distribution with mean $\mu = 6$ and standard deviation $\sigma = 1$, shown in the red curve along the horizontal axis. The non-linear change of variables from x to y is given by:

  $$x = g(y) = \ln y - \ln(1 - y) + 5$$

  And the inverse is a *logistic sigmoid* function given by (shown in the blue curve):

  $$y = g^{-1}(x) = \frac{1}{1 + \exp(-x + 5)}$$

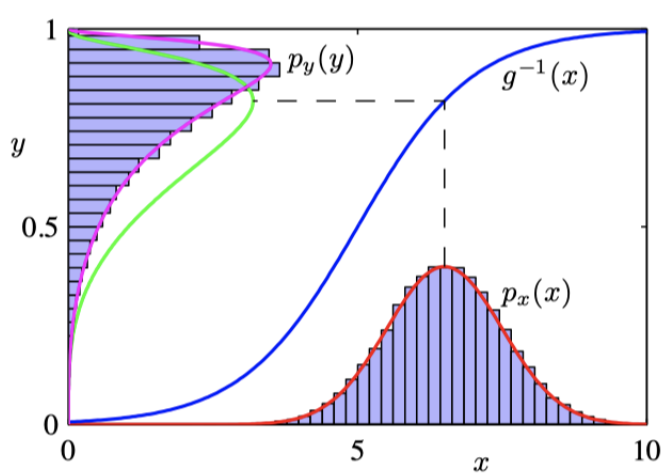

Figure 1: Example of transformation of density and its mode.

If we simply transform $p_x(x)$ as a function of x will lead to the green curve $p_x(g(y))$ in Figure 1. However, the density over y transforms instead according to Equation 1 and is shown by the magenta curve along the vertical axis. Note that this has its mode shifted relative to the mode of the green curve, i.e., the $g^{-1}(\hat{x})$ is not a mode of p_y.

Exercise 3: Variance

Let $f(x)$ be some function in x. Using the definition $\mathrm{Var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right]$ (c.f. [1] 1.38) show that $\mathrm{Var}[f(x)]$ satisfies $\mathrm{Var}[f] = \mathbb{E}\left[f(x)^2\right] - \mathbb{E}[f(x)]^2$

Solution:

$$
\begin{aligned}
\mathrm{Var}[f] &= \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] \\
&= \mathbb{E}\left[f^2(x) + (\mathbb{E}[f(x)])^2 - 2f(x)\,\mathbb{E}[f(x)]\right] \\
&= \mathbb{E}\left[f^2(x)\right] + \mathbb{E}\left[(\mathbb{E}[f(x)])^2\right] - \mathbb{E}\left[2f(x)\,\mathbb{E}[f(x)]\right] && \text{(Linearity of expectation)} \\
&= \mathbb{E}\left[f^2(x)\right] + (\mathbb{E}[f(x)])^2 - \mathbb{E}[f(x)] \cdot \mathbb{E}[2f(x)] \\
&= \mathbb{E}\left[f^2(x)\right] + (\mathbb{E}[f(x)])^2 - 2\,\mathbb{E}[f(x)]^2 \\
&= \mathbb{E}\left[f^2(x)\right] - \mathbb{E}[f(x)]^2
\end{aligned}
$$

Exercise 4: Covariance

Show that if two variables x and y are independent, then their covariance is zero.

Solution:

$$
\begin{aligned}
\mathrm{Cov}(x, y) &= \mathbb{E}[xy] - \mathbb{E}[x]\,\mathbb{E}[y] \\
\mathbb{E}[xy] &= \int\int p(x, y) xy \,\mathrm{d}x\mathrm{d}y \\
&= \int\int p(x)p(y) xy \,\mathrm{d}x\mathrm{d}y && \text{(Independence)} \\
&= \int p(y) y \left(\int p(x) x \,\mathrm{d}x\right) \mathrm{d}y \\
&= \mathbb{E}[x] \int p(y) y \mathrm{d}y \\
&= \mathbb{E}[x]\,\mathbb{E}[y]
\end{aligned}
$$

Thus, $\mathrm{Cov}(x, y) = 0$

Exercise 5: Normal Mode

Recall the definition of the univariate Gaussian distribution

$$
\mathcal{N}\left(x|\mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{2}
$$

and the definition of the multivariate (D-dimensional) Gaussian distribution

$$
\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{3}
$$

i) Show that the mode (i.e. the maximum) of the Gaussian distribution (Equation 2) is given by μ.

ii) Show that the mode of the multivariate Gaussian (Equation 3) is given by $\boldsymbol{\mu}$.

Solution:

i) To find the mode, we need to compute the derivative of $\mathcal{N}\left(x|\mu,\sigma^2\right)$ w.r.t. x (first order condition):

$$\begin{aligned}\frac{\mathrm{d}\mathcal{N}\left(x|\mu,\sigma^2\right)}{\mathrm{d}x} &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\left(-\frac{2(x-\mu)}{2\sigma^2}\right) \qquad \text{(Chain rule)}\\ &= -\mathcal{N}\left(x|\mu,\sigma^2\right)\frac{x-\mu}{\sigma^2}\end{aligned}$$

The mode $\hat{x}$ should make the derivative equal to 0:

$$-\mathcal{N}\left(\hat{x}|\mu,\sigma^2\right)\frac{\hat{x}-\mu}{\sigma^2} = 0$$

As $\sigma^2 > 0$ and $\mathcal{N}\left(\hat{x}|\mu,\sigma^2\right) > 0$ (density is non-negative everywhere and should be > 0 at its mode) $\implies$ $\hat{x}-\mu$ should be 0 to make the derivative $= 0$.

Verifying the second order condition, we have that the second order derivative at $\hat{x} = \mu$ is indeed < 0.

Hence, the mode of the univariate Gaussian distribution is given by $\hat{x} = \mu$.

ii) Taking the derivative of $\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}\right)$ w.r.t. $\mathbf{x}$:

$$\begin{aligned}\frac{\partial\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}\right)}{\partial\mathbf{x}} &= \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)\frac{\partial}{\partial\mathbf{x}}\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \qquad \text{(Chain rule)}\\ &= -\mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}\right)\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\end{aligned}$$

Since
$$\begin{aligned}\frac{\partial}{\partial\mathbf{x}}\left((\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) &= \frac{\partial}{\partial\mathbf{x}}\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}-\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x}-\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\\ &= \frac{\partial}{\partial\mathbf{x}}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{\partial}{\partial\mathbf{x}}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{\partial}{\partial\mathbf{x}}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\\ &= \boldsymbol{\Sigma}^{-1}\mathbf{x}+(\boldsymbol{\Sigma}^{-1})^{\mathrm{T}}\mathbf{x}-(\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1})^{\mathrm{T}}-\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \qquad \text{(Matrix derivatives)}\\ &= 2\boldsymbol{\Sigma}^{-1}\mathbf{x}-2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \qquad \text{(Covariance matrix is symmetric)}\end{aligned}$$

The mode $\hat{\mathbf{x}}$ should make the derivative equal to 0:

$$-\mathcal{N}\left(\hat{\mathbf{x}}|\boldsymbol{\mu},\boldsymbol{\Sigma}\right)\boldsymbol{\Sigma}^{-1}(\hat{\mathbf{x}}-\boldsymbol{\mu}) = 0 \qquad (*)$$

Using the fact that $\mathcal{N}\left(\hat{\mathbf{x}}|\boldsymbol{\mu},\boldsymbol{\Sigma}\right) > 0$ and left-multiplying both side of $(*)$ by $\boldsymbol{\Sigma}$ leads to $\hat{\mathbf{x}} = \boldsymbol{\mu}$.

Verifying the second order condition, we have that the Hessian at $\hat{\mathbf{x}} = \boldsymbol{\mu}$ is indeed negative semi-definite, as $\boldsymbol{\Sigma}$ is positive semi-definite.

Hence, the mode is given by $\hat{\mathbf{x}} = \boldsymbol{\mu}$.

Exercise 6: Independence

Suppose that the two variables x and z are statistically independent.

i) Show that the mean satisfies $\mathbb{E}[x+z] = \mathbb{E}[x] + \mathbb{E}[z]$.

ii) Show that the variance satisfies $\mathrm{Var}[x+z] = \mathrm{Var}[x] + \mathrm{Var}[z]$.

Solution:

We show below the derivation for continuous variables. For discrete variables the integrals are replaced by summations, and the same results are again obtained.

i)

$$
\begin{aligned}
\mathbb{E}\left[x+z\right] &= \int\int (x+z)\cdot p(x,z)\mathrm{d}x\mathrm{d}z \\
&= \int\int x\cdot p(x,z)\mathrm{d}x\mathrm{d}z + \int\int z\cdot p(x,z)\mathrm{d}x\mathrm{d}z \\
&= \int x \int p(x,z)\mathrm{d}z\mathrm{d}x + \int z\int p(x,z)\mathrm{d}x\mathrm{d}z && \text{(Rearrange order of double integrals)} \\
&= \int xp(x)\mathrm{d}x + \int zp(z)\mathrm{d}z && \text{(Sum rule)} \\
&= \mathbb{E}\left[x\right] + \mathbb{E}\left[y\right]
\end{aligned}
$$

We just prove the linearity of expectation. Note that this property also holds, no matter whether the variables are independent or not.

ii)

$$
\begin{aligned}
\mathrm{Var}[x+z] &= \mathbb{E}\left[(x+z-\mathbb{E}\left[x+z\right])^2\right] \\
&= \mathbb{E}\left[((x-\mathbb{E}\left[x\right]) + (z-\mathbb{E}\left[z\right]))^2\right] && \text{(Linearity of expectation, and rearrange)} \\
&= \mathbb{E}\left[(x-\mathbb{E}\left[x\right])^2 + (z-\mathbb{E}\left[y\right])^2 - 2(x-\mathbb{E}\left[x\right])(z-\mathbb{E}\left[z\right])\right] \\
&= \mathbb{E}\left[(x-\mathbb{E}\left[x\right])^2\right] + \mathbb{E}\left[(z-\mathbb{E}\left[y\right])^2\right] - 2\mathrm{Cov}(x,z) && \text{(Linearity of expectation)} \\
&= \mathbb{E}\left[(x-\mathbb{E}\left[x\right])^2\right] + \mathbb{E}\left[(z-\mathbb{E}\left[y\right])^2\right] && (\mathrm{Cov}(x,z)=0\text{, proved in Exercise 4)} \\
&= \mathrm{Var}(x) + \mathrm{Var}(z)
\end{aligned}
$$

References

[1] C. M. Bishop. *Pattern recognition and machine learning.* Springer, 2006.

[2] R. O. Duda, P. E. Hart, and D. G. Stork. Pattern classification. *A Wiley-Interscience Publication*, 2001.