— *Solution notes* —

First practicals in Machine learning 1 – 2023 – Paper 1

1 Calculus (September)

Find the derivatives of the following functions with respect to x.

(a) $\sigma(x) = \frac{1}{1+e^{-x}}$ (the standard logistic function or "sigmoid function").

Answer:

To calculate the derivative of the *sigmoid function* we need to apply the chain rule:

$$\frac{d}{dx}(1+e^{-x})^{-1} = (1+e^{-x})^{-2}e^{-x} = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{e^{-x}}{(1+e^{-x})}\sigma(x)$$
$$= (\frac{e^{-x}+1}{e^{-x}+1} - \frac{1}{e^{-x}+1})\sigma(x) = \sigma(x)(1-\sigma(x))$$

(b) $\max\{0, x\}$ ("Rectified Linear Unit" or ReLu that is important in Deep Learning).

Answer:

$$\frac{d \max\{0,x\}}{dx} = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x < 0 \\ \text{Undefined} & \text{if } x = 0 \end{cases}$$

(c) What is the shape of the following derivative: $\frac{df(x)}{dx}$ $f : \mathbb{R} \to \mathbb{R}$, $x \in \mathbb{R}$

Answer: $\frac{df(x)}{dx} \in \mathbb{R}$

(d) What is the shape of the following derivative:$\frac{df(\boldsymbol{x})}{d\boldsymbol{x}}$ with $f : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{x} \in \mathbb{R}^n$

Answer:

Given our convention the derivative will be a row vector, in particular:

$\frac{df(\mathbf{x})}{d\mathbf{x}} \in \mathbb{R}^{1\times n}$

(e) What is the shape of the following derivative: $\frac{d\boldsymbol{f}(\boldsymbol{x})}{d\boldsymbol{x}}$ with $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$, $\boldsymbol{x} \in \mathbb{R}^n$

Answer: $\frac{d\mathbf{f}(\mathbf{x})}{d\boldsymbol{x}} \in \mathbb{R}^{m\times n}$

(f) Calculate the following derivative $\frac{df(\mathbf{x})}{d\mathbf{x}}$ with $f(\mathbf{x}) = 2\exp(x_2 - \ln(x_1^{-1}) - \sin(x_3 x_1^2)), \mathbf{x} \in \mathbb{R}^3$.

Answer:

Given our convention the derivative will be a row vector, in particular:

$$\frac{d}{d\mathbf{x}} = \begin{bmatrix} \frac{df(\mathbf{x})}{dx_1} \\ \frac{df(\mathbf{x})}{dx_2} \\ \frac{df(\mathbf{x})}{dx_3} \end{bmatrix}^T$$

$$\frac{df(\mathbf{x})}{dx_1} = f(\mathbf{x}) \left(\frac{1}{x} - 2x_3 x_1 \cos(x_3 x_1^2) \right)$$

$$\frac{df(\mathbf{x})}{dx_2} = f(\mathbf{x})$$

$$\frac{df(\mathbf{x})}{dx_3} = -f(\mathbf{x}) \cos(x_3 x_1^2) x_1^2$$

(*g*) $\nabla_y h$ with $h = g(f(y))$ where $g(\mathbf{x}) = x_1^3 + \exp(x_2)$ and $\mathbf{x} := \mathbf{f}(y) = [y\sin(y), y\cos(y)]^T$. First show your understanding of the chain rule before plugging in the actual derivatives.

Answer:

By applying the chain rule we can write:

$$\begin{aligned} \nabla_y h &= \frac{dg(\mathbf{x})}{d\mathbf{x}} \frac{d\mathbf{x}}{dy} \\ &= \begin{bmatrix} 3x_1^2 & \exp x_2 \end{bmatrix} \begin{bmatrix} \sin(y) + y\cos(y) \\ \cos(y) - y\sin(y) \end{bmatrix} \end{aligned}$$

(*h*) We now assume that $\mathbf{x} := \mathbf{f}(y, z) = [y\sin(y) + z, y\cos(y) + z^2]^T$. Provide $\nabla_{y,z} h$. Hint: To determine the correct shape of $\nabla_{y,z} h$, view the input pair y and z as a vector $[y, z]^T$.

Answer:

Firstly, lets denote $\mathbf{a} = [y, z]^T$. Then, by applying the chain rule, as usual:

$$\frac{dh}{d\mathbf{a}} = \frac{dg(\mathbf{x})}{d\mathbf{x}} \frac{d\mathbf{x}}{d\mathbf{a}}$$

The first part will be $\mathbb{R}^{1\times 2}$ and the second will be $\mathbb{R}^{2\times 2}$ resulting in a final shape of 1×2.

$$\frac{dg(\mathbf{x})}{d\mathbf{x}} = \begin{bmatrix} 3x_1^2 & \exp(x_2) \end{bmatrix}$$

$$\frac{d\mathbf{x}}{d\mathbf{a}} = \begin{bmatrix} \sin(y) + y\cos(y) & 1 \\ \cos(y) - y\sin(y) & 2z \end{bmatrix}$$

First practicals in Machine learning 1 – 2022 – Paper 1

2 Multivariate calculus (September)

The following questions are good practice in manipulating vectors and matrices and they are essential for solving for posterior distributions. Given the following expression:

$$(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})+(\boldsymbol{\mu}-\boldsymbol{\mu}_{\mathbf{0}})^T\mathbf{S}^{-\mathbf{1}}(\boldsymbol{\mu}-\boldsymbol{\mu}_{\mathbf{0}})$$

where $\mathbf{x}$, $\boldsymbol{\mu}$, $\boldsymbol{\mu}_0$ are vectors and $\boldsymbol{\Sigma}^{-1}$ and $\mathbf{S}^{-1}$ are "symmetric", "positive semi-definite" "invertible" matrices.

Answer the following questions:

(*a*) Expand the expression and gather terms.

Answer:

Firstly, it will be beneficial for us to identify the dimensionality of each of the terms in the expression. Hence, $\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\mu}_{\mathbf{0}} \in \mathbb{R}^n$ and $\boldsymbol{\Sigma}, \mathbf{S} \in \mathbb{R}^{n\times n}$.

Then, due to the distributive property, we can expand the expression:

$$\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}+\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}+\boldsymbol{\mu}^T\mathbf{S}^{-1}\boldsymbol{\mu}+\boldsymbol{\mu}_0^T\mathbf{S}^{-1}\boldsymbol{\mu}_0-2\boldsymbol{\mu}^T\mathbf{S}^{-1}\boldsymbol{\mu}_0$$

(*b*) Collect all the terms that depend on $\boldsymbol{\mu}$ and those that do not.

Answer:

Note that if we take the transpose of a scalar it is the same value, and therefore, we can have:

$$\mathbf{x}^T\mathbf{A}\mathbf{y}=(\mathbf{x}^T\mathbf{A}\mathbf{y})^T=(\mathbf{y}^T\mathbf{A}^T\mathbf{x})$$

.

Then if matrix $\mathbf{A}^T$ is symmetric, we have $\mathbf{A}^T=\mathbf{A}$.

By taking into account the previous two points we can have:

$$\mu^T(\boldsymbol{\Sigma}^{-1}+\mathbf{S}^{-1})\boldsymbol{\mu}-2\boldsymbol{\mu}^T(\boldsymbol{\Sigma}^{-1}\mathbf{x}+\mathbf{S}^{-1}\mu_0)+\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}+\mu_0^T\mathbf{S}^{-1}\mu_0$$

(*c*) Take the derivative with respect to $\boldsymbol{\mu}$, set to 0, and solve for $\boldsymbol{\mu}$.

Answer:

$$f(\boldsymbol{\mu})=\boldsymbol{\mu}^T(\boldsymbol{\Sigma}^{-1}+\mathbf{S}^{-1})\boldsymbol{\mu}-2\mu^T(\boldsymbol{\Sigma}^{-1}\mathbf{x}+\mathbf{S}^{-1}\mu_0)+\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}+\mu_0^T\boldsymbol{\Sigma}^{-1}\mu_0$$

In the previous expression, only the first two terms depend on μ. Thus, you need to calculate the derivatives for the first two terms only.

The first term has a quadratic form similar to $f = \mathbf{x}^T\mathbf{A}\mathbf{x} \in \mathbb{R}$. We need to calculate the derivative for this term using *index notation*. Firstly, we can write f as:

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \sum_{i=1}^{n} x_i \sum_{j=1}^{n} A_{ij}x_j$$

Then we need to calculate $\frac{\partial f}{\partial x_k}$:

$$\frac{\partial f}{\partial x_k} = \frac{\partial}{\partial x_k}\sum_{i=1}^{n} x_i \sum_{j=1}^{n} A_{ij}x_j$$

Next step is to apply the product rule, and therefore we can write:

$$\frac{\partial f}{\partial x_k} = \sum_{i=1}^{n} \frac{\partial x_i}{\partial x_k} \sum_{j=1}^{n} A_{ij}x_j + \sum_{i=1}^{n} x_i \sum_{j=1}^{n} A_{ij}\frac{\partial x_j}{\partial x_k}$$

Since $\frac{\partial x_i}{\partial x_k} = 1$ when i=k and zero otherwise, for clarity, we can make use of kronecker delta:

$$= \sum_{i=1}^{n} \delta_{ki} \sum_{j=1}^{n} A_{ij}x_j + \sum_{i=1}^{n} x_i \sum_{j=1}^{n} A_{ij}\delta_{kj}$$

which is also gives the final derivation over $\frac{\partial f}{\partial x_k}$:

$$= \sum_{j=1}^{n} A_{kj}x_j + \sum_{i=1}^{n} x_i A_{ik}$$

Note that the first term is the dot product between vector $\mathbf{x}$ and and the k-th row of matrix $\mathbf{A}$ while the second term is the dot product between vector $\mathbf{x}$ with the k-th column of the matrix $\mathbf{A}$. Now if we arrange all the partial derivatives in a row vector:

$$\begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \dots \\ \frac{\partial f}{\partial x_k} \\ \dots \\ \frac{\partial f}{\partial x_N} \end{bmatrix}^T = \begin{bmatrix} \sum_{j=1}^{n} x_j A_{1j} + \sum_{i=1}^{n} x_i A_{i1} \\ \sum_{j=1}^{n} x_j A_{2j} + \sum_{i=1}^{n} x_i A_{i2} \\ \dots \\ \sum_{j=1}^{n} x_j A_{kj} + \sum_{i=1}^{n} x_i A_{ik} \\ \dots \\ \sum_{j=1}^{n} x_j A_{Nj} + \sum_{i=1}^{n} x_i A_{iN} \end{bmatrix}^T = \mathbf{x}^T\mathbf{A}^T + \mathbf{x}^T\mathbf{A} = \mathbf{x}^T(\mathbf{A} + \mathbf{A}^T) = 2\mathbf{x}^T\mathbf{A}$$

Since $\mathbf{A}^T = \mathbf{A}$. Therefore, in our example, we have:

$$\frac{\partial f(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = 2\boldsymbol{\mu}^T(\boldsymbol{\Sigma}^{-1} + \mathbf{S}^{-1}) - 2(\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{S}^{-1}\mu_0)^T = \mathbf{0}$$

We can now move the second term in the other side of the equation and then transpose in both sides:

$$\Leftrightarrow (\boldsymbol{\Sigma}^{-1} + \mathbf{S}^{-1})\boldsymbol{\mu} = (\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{S}^{-1}\boldsymbol{\mu}_0)$$

$$\Leftrightarrow \boldsymbol{\mu} = (\boldsymbol{\Sigma}^{-1} + \mathbf{S}^{-1})^{-1}(\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{S}^{-1}\boldsymbol{\mu}_0)$$

Note that the summation of two *symmetric semi-definite positive* matrices leads to a *symmetric semi-definite positive* matrix.

First practicals in Machine learning 1 – 2022 – Paper 1

3 Probability theory (September)

Consider the following setting. You are driving down the street at night and suddenly you see a man climbing through a broken window of a jewelry store. Then, he runs away carrying a bag over his shoulder. For many of us, our gut reaction would be to think the man in question is a criminal. Why do we draw this conclusion instead of another scenario? Let's explore this using the methods of Probability Theory.

(*a*) Explain in words: why would many people draw the conclusion that the man in question is a criminal? Try to think in terms of probability (1-2 sentences are sufficient).

Answer: Any reasonable argument would work here.

(*b*) Show, formally (using probability theory), that the probability of us believing the man is a criminal given our observation is based on our beliefs of making this observation when the man is a criminal and making the observation when the man is not a criminal. Define first your variables for the problem and then your answer.

Answer:

$\{C, O\}$ Let us define the event of the observations as O and the event of the man begin a criminal as C. Then $P(C|O)$ is the probability of the man being a criminal given our observation. Using Bayes rule, we find:

$$P(C|O) = \frac{P(O|C)P(C)}{P(O|C)P(C) + P(O|\neg C)P(\neg C)}$$

.

Clearly, our beliefs of making the observation given that the man is either a criminal or not a criminal play a role here.

(*c*) Let's assume one in every 10^5 people is in fact a criminal, the probability of making this observation when the man is not a criminal is $\frac{1}{10^6}$, and that of making this observation when the man is a criminal is 0.8.

Compute the probability of the man being a criminal based on our observations.

Answer: Using the numbers that were just given we have:

$$P(C|O) = \frac{P(O|C)P(C)}{P(O|C)P(C) + P(O|\neg C)P(\neg C)} = .. \approx 0.89$$

.

(*d*) The next morning you learn that a group of kids have smashed multiple store fronts in your neighborhood. How does this change your beliefs, i.e., do you still think the man is a criminal? Explicitly state which belief updates you make and re-compute the probability of the man being a criminal given our observation. *Note, you do not have to do a Bayesian update or justify your belief update mathematically.*

Answer: Any argument and correct computation is considered correct. Since kids smashed the windows in store fronts, it is more likely to see a non-criminal in the situation we observed. Therefore, we could update our belief of an innocent man being in this situation. E.g., $P(O|C) = 0.01$, which results in:

$$P(C|O) = \frac{P(O|C)P(C)}{P(O|C)P(C) + P(O|\neg C)P(\neg C)} = .. \approx 0.0008$$

.

First practicals in Machine learning 1 – 2022 – Paper 1

4 MAP of Binomial (September)

Suppose someone gave you a magic coin that even though it is perfectly flat, it shows heads much more than tails. You toss it 8 times, resulting in an 8-bit binary string.

(*a*) This sequence can be modelled with a binomial distribution. If the coin is unbiased, what will be the value of p?

Answer: If the coin is truly unbiased it will have value of exactly 0.5 and thus we would expect 4 1s and 4 0s.

(*b*) We observe the bitstring 01110110. What is the maximum likelihood estimate of p? First derive the question analytically, then give the numerical solution.

Answer: The likelihood function

$$f(m|n,p) = \binom{n}{m} p^m (1-p)^{n-m}$$

.

Find extremum using the log-likelihood.

$$\begin{aligned}\frac{d}{dp}\ln f(m|n,p) &= \frac{m}{p} - \frac{n-m}{1-p}\\ \Leftrightarrow m(1-p) &= p(n-m)\\ \Leftrightarrow m &= pn\\ \Leftrightarrow p &= \frac{m}{n}\end{aligned}$$

.

Insert numbers

$$p = 5/8 = 0.625$$

.

(*c*) Knowing this, a naive person would accept the coin as biased. You are not naive, and you've flipped enough coins to know that almost all of them are practically unbiased. You express this by having a strong prior. This prior appears in Bayes' rule. It's a beta distribution, which is the conjugate prior of the binomial, resulting in a beta posterior distribution.

Show that the posterior distribution is beta. Combine all proportionality constants (not depending on p), in a constant Z.

Answer:

Define the posterior:

$$f(p|a,b,m,n) = \frac{f(m|n,p)f(p|a,b)}{f(n,a,b)}$$

.

The likelihood is already modeled by the binomial as:

$$f(m|n,p) = \binom{n}{m} p^m (1-p)^{n-m}$$

.

and the prior is modeled by a Beta distribution:

$$f(p|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1} p^{b-1}$$

.

with a, b being the hyperparameters of the distribution. The evidence can be calculated by marginalizing over the variable p:

$$f(n,a,b) = \int_0^1 f(m|n,p) f(p|a,b) dp = \int_0^1 f(m,p) dp$$

.

We know also that $\Gamma(a+1) = a!$, so we can re-write the likelihood as:

$$f(m|n,p) = \binom{n}{m} p^m (1-p)^{n-m} = \frac{\Gamma(n+1)}{\Gamma(n-m+1)\Gamma(m+1)} p^m (1-p)^{n-m}$$

.

Then, we can calculate the joint probability as:

$$f(m,p) = \frac{\Gamma(n+1)}{\Gamma(n-m+1)\Gamma(m+1)} p^m (1-p)^{n-m} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1} (1-p)^{b-1}$$

.

$$\Leftrightarrow f(m,p) = \Omega p^{m+a-1} (1-p)^{n+b-m-1}$$

.

The last expression it looks like a beta with hyperparameters $a^{'} = m+a$ and $b^{'} = n+b-m$, and in that case a Beta distribution is the following:

$$B(a^{'}, b^{'}) = \frac{\Gamma(n+b+a)}{\Gamma(m+a)\Gamma(n+b-m)} p^{m+a-1} (1-p)^{n+b-m-1} = Z p^{m+a-1} (1-p)^{n+b-m-1}$$

.

and I can re-write my joint probability as:

$$f(m,p) = \Omega \frac{1}{Z} Z p^{m+a-1} (1-p)^{n+b-m-1}$$

.

The next step is to calculate the evidence by marginalizing out the p:

$$f(n,a,b) = \int_0^1 f(m,p) dp = \int_0^1 \Omega \frac{1}{Z} Z p^{m+a-1} (1-p)^{n+b-m-1} dp = \frac{\Omega}{Z}$$

.

Hence, finally, if we replace all the terms that we have calculated so far we will have:

$$f(p|a,b,m,n) = \frac{f(m|n,p)f(p|a,b)}{f(n,a,b)} = Zp^{m+a-1}(1-p)^{n+b-m-1} = Beta(a^{'},b^{'})$$

.

(*d*) Incorporating your strong beta prior with $a = 13$ and $b = 13$, what is your skeptical (MAP) estimate of p? Again, first derive analytically, then plug in the numbers.

Answer:

$$\frac{d}{dp}\ln f(|a,b,m,n) = \frac{m+a-1}{p} - \frac{n-m+b-1}{(1-p)} = 0$$

$$\Leftrightarrow (1-p)(m+a-1) = p(n-m+b-1)$$

$$\Leftrightarrow p = \frac{m+a-1}{n+b+a-2}$$

Plug in the numbers.

$$p = \frac{5+13-1}{8+13+13-2} = 0.53125$$

(*e*) How can a and b be interpreted?

After better inspection, you see that the coin has been made using two different kinds of metal. It seems clear now that this is not a regular coin. How does your belief change? What parameters change, and in which direction?

Answer: a and b can be interpreted as extra datapoints (where a denotes the successes and $a + b$ the total observations) that you previously observed. Any other sensible interpretation should also be fine.

After the inspection, your prior becomes less strong, as you have not seen such a coin before, therefore a and b both become smaller, making the distribution less sharply peaked.

(*f*) Could these results have been derived with a Bernoulli distribution, too? If so, explain why and obtain the ML (maximum likelihood) estimate.

Answer: Since we independently conducted 8 trials, and we are only interested in obtaining the parameter p (we saw that when differentiating the binomial coefficient immediately dropped), this could have been modelled with 8 factorised Bernoulli distributions:

$$p(D|p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

The ML estimate:

$$\frac{d}{dp}\log p(D|p) = \sum_{j=1}^{n} xp - (1-x)(1-p) = 0$$

$$\Leftrightarrow p = \frac{1}{n}\sum_{j=1}^{n} x_j = \frac{m}{n}$$

And so we obtain the same result.

First practicals in Machine learning 1 – 2022 – Paper 1

5 Probability theory II (September)

For this question, you will compute the expression for the posterior parameter distribution for a simple data problem. Assume we observe N univariate data points $x_1, x_2, ..., x_N$. Furthermore, we assume that they are generated by a Gaussian distribution with known variance σ^2, but unknown mean μ. Assume a prior Gaussian distribution over the unknown mean, i.e., $p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$. When answering these questions, use $\mathcal{N}(a|b, c^2)$ to indicate a Gaussian (normal) distribution over a with mean b and variance c^2.You do not need to write down the explicit form of a Guassian distribution.

(*a*) Write down the general expression for a posterior distribution, using θ for the parameter, $\mathcal{D}$ for the data. Indicate the prior, likelihood, evidence, and posterior.

Answer:

$$\underbrace{p(\theta|D)}_{\text{posterior}} = \frac{\overbrace{p(\theta)}^{\text{prior}}\ \overbrace{p(D|\theta)}^{\text{likelihood}}}{\underbrace{p(D)}_{\text{evidence}}} = \frac{\overbrace{p(\theta)}^{\text{prior}}\ \overbrace{p(D|\theta)}^{\text{likelihood}}}{\underbrace{\int p(\theta)p(D|\theta)d\theta}_{\text{evidence}}}$$

(*b*) Write the posterior for this particular example. You do not need an analytic solution.

Answer:

$$p(\theta|D) = \frac{\mathcal{N}(\mu|\mu_0, \sigma_0^2) \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2)}{\int \mathcal{N}(\mu|\mu_0, \sigma_0^2) \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2) d\mu}$$