First assignment in Machine learning 1 – 2022 – Paper 1

1 Multivariate Calculus (Recommended timeline: September 14)

In this exercise, you need to compute several gradients. You should simplify your answers as much as possible, *and use index-notation for all your derivations.* Consider $\nabla f_{\mathbf{x}}(\mathbf{x})$ as the same with $\frac{df}{d\mathbf{x}}$). Compute the following:

(*a*) $\nabla_{\mathbf{x}}\sigma(\mathbf{x})$ with $\mathbf{x} \in \mathbb{R}^m$ where σ denotes the Sigmoid function applied element-wise. [1 point]

(*b*) $\frac{d}{d\mathbf{w}}\mathbf{f}$ with $\mathbf{f} = \mathbf{X}\mathbf{w}$ with $\mathbf{X} \in \mathbb{R}^{n\times m}$ and $\mathbf{w} \in \mathbb{R}^m$ [1 point]

(*c*) $\frac{d}{d\mathbf{w}}f$ with $f = \mathbf{w}^T\mathbf{X}\mathbf{w}$ with $\mathbf{X} \in \mathbb{R}^{n\times m}$ and $\mathbf{w} \in \mathbb{R}^m$ [1 point]

(*d*) Compute the following gradient, by assuming $\mathbf{\Sigma^{-1}}$ is symmetric (Hint: For a symmetric matrix $\mathbf{\Sigma}$ it holds that $\Sigma_{ij} = \Sigma_{ji}$), positive semi-definite and invertible:

$\frac{d}{d\mathbf{\Sigma}^{-1}}(\mathbf{x}-\mathbf{A}\mathbf{s})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{A}\mathbf{s})$ with $(\mathbf{x}-\mathbf{A}\mathbf{s}) \in \mathbb{R}^m$ and $\mathbf{\Sigma}^{-1} \in \mathbb{R}^{m\times m}$ [1 point]

(*e*) $\frac{d}{d\mathbf{x}}\varsigma(\mathbf{x}), \mathbf{x} \in \mathbb{R}^n$ where $\varsigma(\mathbf{x})_i = \frac{\exp x_i}{\sum_{j=1}^n \exp x_j}$. Try to write it in matrix form making use of the diag function. [2 points]

Hint: Try to write the seperate terms in terms of p(X=k) and use the normalization property. [2 points]

First assignment in Machine learning 1 – 2022 – Paper 1

2 Full analysis of a distribution: Poisson distribution (Recommended timeline: September 21)

The Poisson process is a model for series of discrete events where an average time between events is known, but the exact time at which they occur is random. It is also assumed that the process is memoryless or *Markovian*, i.e. the occurrence of a new event is independent of the previous events.

There are many processes that are modeled this way. For example, imagine a fast-food restaurant in which a new customer enters on average every minute. The arrival time of the next customer is random and does not depend on the arrival time of the previous customer. Thus, the Poisson distribution is a suitable choice for modeling the number of customers arriving in some interval of time. There are many real-world examples where Poisson distribution is used: a daily number of visitors to a website, a number of patients arriving to the hospital every hour, a yearly number of meteors hitting the Earth, and many more. Although these examples are associated with the time domain, there are many examples of Poisson processes associated with space (number of events per area). For example, a number of chewing gums per m^2 on the street follows a Poisson distribution, as well as the number trees within an acre in a forest.

Formally, a discrete random variable X (number of occurrences) is said to have a Poisson distribution, with a parameter $\lambda > 0$, if the probability of k occurrences is given by:

$$p(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

The aim of this exercise is to familiarize you with analyzing arbitrary distributions. Note that by no means the following questions are the only things you might want to know about a distribution, but rather serve as a starting point for further research. Using these insights, answer the following questions:

Useful identities:

- Taylor expansion of the Exponential function:

$$e^x = \sum_{j=0}^{\infty} \frac{x^j}{j!}$$

(a) Show that the Poisson distribution is well-defined (non-negative probabilities) and is normalized (probabilities sum up to 1).

[1 point]

(*b*) Calculate the mean and the variance of the Poisson distribution.

Hint: Try to write the separate terms in terms of $p(X = k)$ and use the normalization property.

[2 points]

(*c*) Assume that you have measured N occurrences of an event always measured in fixed time interval Δt (a standard assumption for Poisson processes), and you have the following measurements $(k_1, \ldots, k_N)$. Write down the log-likelihood under the i.i.d assumption, and assuming that events are Poisson processes.
[1 point]

(*d*) One common method to estimate the parameters of the assumed probability distribution is called maximum likelihood optimization. In this approach, we wish to find the parameters of the distribution (in our case λ) which will maximize the likelihood function. Find the maximum likelihood estimator λ_{ML} for the likelihood function calculated in part c). [2 points]

(*e*) Imagine that you have obtained an estimate λ_{ML} when measuring the number of customers entering a store every minute. How would you obtain an average number of customers entering a store every hour? [1 point]

(*f*) In general it is difficult to find a closed form expression for the posterior distribution of our model parameters because of the integral in the evidence. Rather than finding the full posterior distribution, we can find a point estimate of the model parameters. A common point estimate is the maximum a posteriori estimation, or the MAP, which estimates the model parameters as the mode of the posterior distribution, i.e.

$$\lambda_{\mathrm{MAP}} = \arg\max_{\lambda} p(\lambda \mid k).$$

Assume that the prior for the parameter λ is given by the Gamma distribution with hyperparameters α_1 and α_2:

$$p(\lambda|\alpha_1, \alpha_2) = \frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)} \lambda^{\alpha_1 - 1} e^{-\alpha_2 \lambda},$$

where Γ denotes the gamma function. Show that we can find λ_{MAP} by optimizing

$$\lambda_{\mathrm{MAP}} = \arg\max_{\lambda} \; ((\sum_{i=1}^{N} k_i) + \alpha_1) \log \lambda - (N - \alpha_2)\lambda.$$

Hint: Show first that $\lambda_{\mathrm{MAP}} = \arg\max_{\lambda} \log p(k \mid \lambda) + \log p(\lambda)$. [2 points]

(*g*) Find the MAP estimator λ_{MAP}. [1 point]

In the case of a Poisson distribution with a Gamma prior, the resulting posterior distribution can be derived analytically. The resulting distribution is a Gamma distribution $\text{Gamma}(\alpha'_1, \alpha'_2)$.

(*h*) **BONUS:** Show that the posterior distribution is indeed a Gamma distribution.

Hint: The resulting distribution follows $\alpha'_1 = (\sum_{i=1}^{N} k_i) + \alpha_1$ and $\alpha'_2 = N + \alpha_2$. [2 points]

3 General Multiple Outputs Linear Regression (Recommended timeline: September 21)

So far, all linear regression models assumed that the target t is a single target. In a more general case, however, we may wish to predict multiple targets $\mathbf{t}$.

One possibility is to perform an independent linear regression for each component of the target vector $\mathbf{t}$ by introducing a different set of basis functions for each component. In other words, if the target is a K-dimensional vector, we would perform a separate linear regression for each component t_i, $i \in 1, \ldots, K$ of the target vector $\mathbf{t}$.

The other more common approach is to use the same set of basis function to model the target vector directly in the following form:

$$\mathbf{y}(\mathbf{x}, \mathbf{W}) = \mathbf{W}^{\mathrm{T}} \phi(\mathbf{x}),$$

where $\mathbf{y}$ is the model prediction, $\mathbf{x}$ is a M-dimensional input vector, and $\mathbf{W}$ is matrix of parameters. $\phi(\mathbf{x})$ is an M-dimensional vector with elements $\phi_j(\mathbf{x})$, and $\phi_0(\mathbf{x}) = 1$ as usual. Assume that the conditional distribution of the target vector to be an Gaussian of the form:

$$p\left(\mathbf{t} \middle| \mathbf{W}, \boldsymbol{\Sigma}\right) = \mathcal{N}\left(\mathbf{t} \middle| \mathbf{y}(\mathbf{x}, \mathbf{W}), \boldsymbol{\Sigma}\right),$$

where $\boldsymbol{\Sigma}$ is the covariance matrix, and $\mathbf{y}(\mathbf{x}, \mathbf{W}) = \mathbf{W}^{\mathrm{T}} \phi(\mathbf{x})$. Assume that we have N observations $\mathbf{t_1}, \ldots, \mathbf{t_N}$, which can be combined into a matrix $\mathbf{T}$ of size $N \times K$, such that the n^{th} row is given by $\mathbf{t}_n^{\mathrm{T}}$.

(a) What are the dimensions of the parameter matrix $\mathbf{W}$? [1 point]

(b) Write down the log-likehood. [1 point]

(c) Find the maximum likelihood solution $\mathbf{W}_{\mathrm{ML}}$ in the terms of feature matrix $\boldsymbol{\Phi}$ and the target matrix $\mathbf{T}$, and show that it is independent of the covariance matrix $\boldsymbol{\Sigma}$. [2 points]

(d) **BONUS:** Show that the maximum likelihood solution for $\boldsymbol{\Sigma}$ is given by:

$$\boldsymbol{\Sigma} = \frac{1}{N} \sum_{n=1}^{N} \left(\mathbf{t}_n - \mathbf{W}_{\mathrm{ML}}^{\mathrm{T}} \phi(\mathbf{x}_n)\right) \left(\mathbf{t}_n - \mathbf{W}_{\mathrm{ML}}^{\mathrm{T}} \phi(\mathbf{x}_n)\right)^{\mathrm{T}}$$

[2 points]

Hint: Substitute $\boldsymbol{\Omega} := \boldsymbol{\Sigma}^{-1}$ and differentiate with respect to $\boldsymbol{\Omega}$. Moreover, make use of the identities $\frac{d}{d\mathbf{A}} \log |\mathbf{A}| = (\mathbf{A^{-1}})^{\mathbf{T}}$ and $\frac{d}{d\mathbf{X}} \mathbf{a^T X^T b} = \mathbf{b a^T}$.

4 Bayesian Linear Regression (Recommended timeline: September 21)

Consider a linear basis function model (such as in Bishop 3.1), and suppose that we have already observed N data points so that the posterior distribution over $\boldsymbol{w}$ for the observed data (N) is given by:

$$p(\boldsymbol{w}|\boldsymbol{t}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N) \ ,$$

where

$$\boldsymbol{\mu}_N = \boldsymbol{\Sigma}_N(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \beta\boldsymbol{\Phi}_N\boldsymbol{t}_N) \ ,$$

and

$$\boldsymbol{\Sigma}_N^{-1} = \boldsymbol{\Sigma}_0^{-1} + \beta\boldsymbol{\Phi}_N^T\boldsymbol{\Phi}_N \ .$$

When a new data-point,$(\boldsymbol{x}_{N+1}, t_{N+1})$, is observed, this posterior can be regarded as the prior for the next observation. By considering this additional data-point, show that the resulting posterior distribution is equal to:

$$p(\boldsymbol{w}|\boldsymbol{t}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_{N+1}, \boldsymbol{\Sigma}_{N+1}) \ ,$$

with

$$\boldsymbol{\mu}_{N+1} = \boldsymbol{\Sigma}_{N+1}(\boldsymbol{\Sigma}_N^{-1}\boldsymbol{\mu}_N + \beta\boldsymbol{\phi}_{N+1}^T t_{N+1}) \ ,$$

and

$$\boldsymbol{\Sigma}_{N+1}^{-1} = \boldsymbol{\Sigma}_N^{-1} + \beta\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}_{N+1}^T \ .$$

To do this, answer the following questions:

(a) Which prior distribution $p(\mathbf{w}|\mathbf{S}_0, \mathbf{m}_0)$ has been used to obtain the reported posterior? Write your answer as a function of $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\mu}_0$. [1 point]

Hint: what happens to the posterior distribution when no data is observed?

(b) What happens to the posterior update rule when the precision β approaches 0? Explain with your own words. [1 point]

(c) **BONUS**: When a new example, $(\boldsymbol{\phi}_{N+1}, t_{N+1})$, is observed, the posterior $p(\mathbf{w}|\boldsymbol{\Phi}_N, \mathbf{t}_N, \boldsymbol{\Sigma}_0, \boldsymbol{\mu}_0, \beta)$ can be regarded as the prior for the next observation. Write down the expression for the updated posterior after observing the new data-point $p(\mathbf{w}|\boldsymbol{\Phi}_{N+1}, \mathbf{t}_{N+1}, \boldsymbol{\Sigma}_0, \boldsymbol{\mu}_0, \beta)$ as a function of the old posterior, likelihood of the new observation and its evidence. [1 point]

(d) **BONUS**: By expanding the expression reported in the previous answer, show that the updated posterior distribution is equal to:

$$p(\mathbf{w}|\boldsymbol{\Phi}_{N+1}, \mathbf{t}_{N+1}, \boldsymbol{\Sigma}_0, \boldsymbol{\mu}_0, \beta) = \mathcal{N}(\mathbf{w}|\boldsymbol{\mu}_{N+1}, \boldsymbol{\Sigma}_{N+1}) \ ,$$

with

$$\boldsymbol{\mu}_{N+1} = \boldsymbol{\Sigma}_{N+1}(\boldsymbol{\Sigma}_N^{-1}\boldsymbol{\mu}_N + \beta\boldsymbol{\phi}_{N+1}^T t_{N+1}) \ ,$$
$$\boldsymbol{\Sigma}_{N+1}^{-1} = \boldsymbol{\Sigma}_N^{-1} + \beta\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}_{N+1}^T \ .$$

Hint: focus on the terms inside the exponential; you can discard normalization terms which do not depend on $\mathbf{w}$ during your derivation to identify mean and covariance of the updated posterior.

[1 point]