2 Multivariable Calculus

Question 2.1

The following questions are good practice in manipulating vectors and matrices. Compute the following gradients, assuming $\boldsymbol{\Sigma}^{-1}$ is symmetric, positive semidefinite and invertible. Simplify your answers as much as possible.

(a) $\nabla_{\boldsymbol{\mu}}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$

Answer:

$$\begin{aligned}\nabla_{\boldsymbol{\mu}}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) =& \nabla_{\boldsymbol{\mu}}(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}-\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1})(\boldsymbol{x}-\boldsymbol{\mu})\\ =& \nabla_{\boldsymbol{\mu}}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}-\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}+\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\\ =& -\boldsymbol{\Sigma}^{-1}\boldsymbol{x}-\boldsymbol{\Sigma}^{-1}\boldsymbol{x}+2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\\ =& 2\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{x})\end{aligned}$$

(b) $\nabla_{\boldsymbol{q}}-\boldsymbol{p}^T log(\boldsymbol{q})$, where log(.) is applied element wise.

Answer:

$$\begin{aligned}\nabla_{\boldsymbol{q}}-\boldsymbol{p}^T log(\boldsymbol{q}) =& -\nabla_{\boldsymbol{q}}\sum_{i=1}^{N} p_i log(q_i)\\ =& -\left[\tfrac{p_1}{q_1}\dots\tfrac{p_n}{q_n}\right]\end{aligned}$$

(c) $\nabla_{\boldsymbol{Wf}}$, where $\boldsymbol{f}=\boldsymbol{W}\boldsymbol{x}, \boldsymbol{W}\in\mathbb{R}^{2\times3}$, and $\boldsymbol{x}\in\mathbb{R}^3$. Follow Example 5.11 of the book mathematics for machine learning to solve this

Answer: We start by determining the dimension of the gradient as

$$\frac{d\boldsymbol{f}}{d\boldsymbol{A}}\in\mathbb{R}^{M\times(M\times N)} \implies \frac{d\boldsymbol{f}}{d\boldsymbol{A}}\in\mathbb{R}^{2\times(2\times3)}$$

The gradient will be the collection of the partial derivatives

$$\frac{d\boldsymbol{f}}{d\boldsymbol{W}}=\begin{bmatrix}\frac{\partial f_1}{\partial \boldsymbol{W}}\\ \\ \frac{\partial f_2}{\partial \boldsymbol{A}}\end{bmatrix}$$

$$f_i = \sum_{j=1}^{3} W_{ij} x_j \qquad i = 1, 2$$

and the partial derivative is

$$\frac{\partial f_i}{\partial W_{iq}} = x_q$$

We compute the partial derivatives of f_i with respect to a row of $\boldsymbol{W}$ which is given as

$$\frac{\partial f_i}{\partial W_{i,:}} = \boldsymbol{x}^T \in \mathbb{R}^{1\times(1\times 3)}$$

and

$$\frac{\partial f_i}{\partial W_{k\neq i}} = \mathbf{0}^T \in \mathbb{R}^{1\times(1\times 3)}$$

giving us

$$\frac{df}{d\boldsymbol{W}} = \begin{bmatrix} \begin{bmatrix} x_1 & x_2 & x_3 \\ 0 & 0 & 0 \end{bmatrix} \\ \begin{bmatrix} 0 & 0 & 0 \\ x_1 & x_2 & x_3 \end{bmatrix} \end{bmatrix}$$

(d) $\nabla_{\boldsymbol{W}} f$, where $f = (\boldsymbol{\mu} - \boldsymbol{W}\boldsymbol{x})^T \boldsymbol{\Sigma^{-1}} (\boldsymbol{\mu} - \boldsymbol{W}\boldsymbol{x})$ where $\boldsymbol{W} \in \mathbb{R}^{M\times K}$

Answer:

$$\begin{aligned} \nabla_{\boldsymbol{W}} f(\boldsymbol{\mu} - \boldsymbol{W}\boldsymbol{x})^T \boldsymbol{\Sigma^{-1}} (\boldsymbol{\mu} - \boldsymbol{W}\boldsymbol{x}) &= \nabla_{\boldsymbol{W}} f[\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} - \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{W}\boldsymbol{x} - \boldsymbol{W}^T \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} + \boldsymbol{W}^T \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{W}\boldsymbol{x}] \\ &= \nabla_{\boldsymbol{W}} f[\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} - 2\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{W}\boldsymbol{x} + \boldsymbol{W}^T \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{W}\boldsymbol{x}] \\ &= -2\boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}\boldsymbol{x}^T + 2\boldsymbol{\Sigma}^{-1} \boldsymbol{W}\boldsymbol{x}\boldsymbol{x}^T \\ &= -2\boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{W}\boldsymbol{x})\boldsymbol{x}^T \end{aligned}$$

3 Probability theory

Question 3.1

A little warmup. This question is based on one of the early chapters in Probability Theory: The Logic of Science by E.T. Jaynes. 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? Lets 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. [13 sentences is sufficient]

Answer: If we base our gut reaction on previous knowledge gained from similar circumstances through various scenarios that are similar in nature, it is clear that the amount of times the person was a criminal far exceeds the amount of times the person has not been a criminal. Given that information we can deduct that the likelihood of the person being a criminal far exceeds the likelihood that he isn't. Because of this we see that the expected result is him being a criminal and that is the conclusion our gut reaction jumps towards.

(b) Show, formally, 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.

Lets 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.

Answer:

Let X denote men and Y denote our assumption on whether or not a man is a criminal. We are given that on in every 10^5 men are criminals, based on that we can derive that

$$P(X = Criminal) = \frac{1}{10^5} \qquad P(X = Innocent) = 1 - \frac{1}{10^5}$$

We are also given that when we believe someone is a criminal, there is a $\frac{1}{10^6}$ chance that the man is actually innocent and when we accuse someone of being a criminal, there is a 0.8 chance that we are correct. Based on that we can derive the following formulas

$$P(Y = Criminal|X = Innocent) = \frac{1}{10^6} \qquad P(Y = Criminal|X = Criminal) = 0.8$$

which allows us to also derive the following

$$P(Y = Innocent|X = Innocent) = 1 - \frac{1}{10^6} \quad P(Y = Innocent|X = Criminal) = 1 - 0.8$$

We can use the sum rule to calculate the probability of a person being assumed to be a criminal.

$$\begin{aligned} P(Y = Criminal) =& P(Y = C|X = C)P(X = C) + P(Y = C|X = I)P(X = I) \\ =& 0.8 \times \frac{1}{10^5} + \frac{1}{10^6} \times (1 - \frac{1}{10^5}) \\ \approx& 9.0 \times 10^{-6} \end{aligned}$$

And the probability of a person being assumed innocent is

$$P(Y = Innocent) \approx 1 - 9.0 \times 10^{-6}$$

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

Answer:

The equation we are asked to solve is the following

$$P(X = Criminal) = P(X = C|Y = C)P(Y = C) + P(X = C|Y = I)P(Y = I)$$

First we can use Bayes to find the following

$$P(X = C|Y = C) \qquad\qquad P(X = C|Y = I)$$

$$\begin{aligned} P(X = C|Y = C) &= \frac{P(Y = C|X = C)P(X = C)}{P(Y = C)} \\ &= \frac{0.8 \times \frac{1}{10^5}}{9.0 \times 10^{-6}} \\ &\approx 0.889 \end{aligned}$$

$$\begin{aligned} P(X = C|Y = I) &= \frac{P(Y = I|X = C)P(X = C)}{P(Y = I)} \\ &= \frac{0.2 \times \frac{1}{10^5}}{1 - 9.0 \times 10^{-6}} \\ &\approx 2 \times 10^{-6} \end{aligned}$$

Now we substitute the values into the equation and calculate the probability of a person being a criminal.

$$\begin{aligned} P(X = Criminal) =& P(X = C|Y = C)P(Y = C) + P(X = C|Y = I)P(Y = I) \\ &= 0.889 \times 9.0 \times 10^{-6} + 2 \times 10^{-6} \times (1 - 9.0 \times 10^{-6}) \\ &\approx 1 \times 10^{-5} \end{aligned}$$

(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: If we factor in this new information to be included in our set of all previous experience then there could be a slight decrease in the likelihood that this individual is a criminal. Although the likelihood has decreased it is only a small amount and thus he is still likely to be a criminal.

Question 3.2

(a) We are asked to write down the expression for the likelihood of the observed datapoints.

$$p(\mathcal{D}|\boldsymbol{\rho}) = \prod_{k=1}^{K} \rho_k^{m_k}$$

where

$$m_k = \sum_n x_{nk}$$

(b) Based on our observed data we can compute ρ_i as

$$\rho_i = \frac{1}{N} \sum_{j=1}^{N} x_{ji}$$

which gives us

$$\boldsymbol{\rho} = \begin{bmatrix} \frac{1}{5} & 0 & 0 & \frac{1}{5} \end{bmatrix}^T$$

(c) We user Bernoulli distribution to describe the probability of drawing a red card.

The likelihood function for the Bernoulli distribution is given by

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

We can rewrite the expression so that p denotes the probability of a red card being drawn and x_{n1} and x_{n3} being a red card as

$$p(\mathcal{D}|p) = \prod_{n=1}^{N} p^{x_{n1}+x_{n3}}(1-p)^{1-x_{n1}+x_{n3}}$$

and the probability of a card being red is written as

$$p = \frac{1}{N}\sum_{j=1}^{N} x_{j1} + x_{j3}$$

(d) Using the formula we defined to evaluate p we can compute the values of p that most likely generate these observations

$$p = \frac{1}{N}\sum_{j=1}^{N} x_{j1} + x_{j3} = \frac{6}{8} = \frac{3}{4}$$

(e) p can be computed by taking the sum of all ρ_i where ρ_i is a red card, in our specific case it can be represented as

$$p = \rho_1 + \rho_3$$

(f)

$$p(p|\mathcal{D}) = \frac{p(\mathcal{D}|p)p(p)}{p(\mathcal{D})}$$

Where

$p(p|\mathcal{D})$ represents the posterior

$p(\mathcal{D}|p)$ represents the likelihood

$p(p)$ represents the prior

$p(\mathcal{D})$ represents the evidence

(g) We need to use the the equation defined in f to calculate the maximum a posterior

$$p(p|\mathcal{D}) = \frac{p(\mathcal{D}|p)p(p)}{p(\mathcal{D})}$$

where our likelihood is defined as

$$p(\mathcal{D}|p) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

and the prior is defined as

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

We can ignore $p(\mathcal{D}$ since we don't need normalization when finding the maximum a posterior

We fit these equations for our parameters and get that

$$p(p|\mathcal{D}) = (\prod_{n=1}^{N} p^{x_n}(1-p)^{1-x_n})(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1})$$

We can further rewrite the expression for posterior distribution as defined in bishop 2.17 and end up with

$$p(p|m,l,\alpha,\beta) \propto p^{m+\alpha-1}(1-p)^{l+\beta-1}$$

where

$$m = \sum_{n=1}^{N} x_i$$

and

$$l = 1 - \sum_{n=1}^{N} x_i$$

We now find the maximum a posterior by taking the derivative of $log(p(p|m, l, \alpha, \beta)$ with respect to p and solve for p

start by simplifying the equation

$$\begin{aligned} p(p|m, l, \alpha, \beta) =& log(p^{m+\alpha-1}(1-p)^{l+\beta-1}) \\ =& log(p^{m+\alpha-1}) + log((1-p)^{l+\beta-1}) \\ =& (m+\alpha-1)log(p) + (l+\beta-1)log(1-p) \end{aligned}$$

we now take the derivative with respect to p

$$\frac{d}{dp}p(p|m, l, \alpha, \beta) = \frac{m+\alpha-1}{p} - \frac{l+\beta-1}{1-p}$$

now we solve for p

$$\begin{aligned} \frac{m+\alpha-1}{p} - \frac{l+\beta-1}{1-p} &= 0 \\ \frac{m+\alpha-1}{p} &= \frac{l+\beta-1}{1-p} \\ (1-p)(m+\alpha-1) &= p(l+\beta-1) \\ m+\alpha-1-p(m+\alpha-1) &= p(l+\beta-1) \\ m+\alpha-1 &= p(l+\beta-1)+p(m+\alpha-1) \\ m+\alpha-1 &= p(l+\beta+m+\alpha-2) \\ \frac{m+\alpha-1}{l+\beta+m+\alpha-2} &= p \end{aligned}$$

so we have found the equation that solves for the maximum a posterior which is defined as

$$p_{MAP} = \frac{m+\alpha-1}{l+\beta+m+\alpha-2}$$

We should find values for α and β that are equal and high in certainty rather than low based on our knowledge that the prior probability distribution is quite certain. Good values would be proportional to our former belief which would be

$$\alpha = \beta = 26$$