AVD613 - Machine Learning for Signal Processing

Tutorial 1 Solutions

Submitted by

Vatsalya Gupta
SC19B098
B.Tech. ECE VII Semester

Department of Avionics
Indian Institute of Space Science and Technology
Thiruvananthapuram - 695 547
August 2022

Solution 1 - Partial Derivatives

(a) Given $L = \frac{1}{2}(y - f(x))^2$

$$\implies \frac{\partial L}{\partial w} = \frac{1}{2}\frac{\partial (y - f(x))^2}{\partial w} = -(y - f(x))\frac{\partial (f(x))}{\partial w} = \frac{1}{2}(f(x) - y)\frac{\partial (1 + tanh(\frac{wx+b}{2}))}{\partial w}$$

$$= \frac{1}{2}(f(x) - y)\frac{\partial (tanh(\frac{wx+b}{2}))}{\partial w} = \frac{1}{2}(f(x) - y)\frac{\partial \left(\frac{e^{\frac{wx+b}{2}} - e^{-\frac{wx+b}{2}}}{e^{\frac{wx+b}{2}} + e^{-\frac{wx+b}{2}}}\right)}{\partial w} = \frac{1}{2}(f(x) - y)\frac{\partial (\frac{e^{wx+b}-1}{e^{wx+b}+1})}{\partial w}$$

$$= \frac{f(x) - y}{2}\left[\frac{(xe^{wx+b})(e^{wx+b} + 1) - (xe^{wx+b})(e^{wx+b} - 1)}{(e^{wx+b} + 1)^2}\right] = (f(x) - y)\left[\frac{xe^{wx+b}}{(e^{wx+b} + 1)^2}\right] \quad (1)$$

$$\implies \frac{\partial L}{\partial b} = \frac{1}{2}\frac{\partial (y - f(x))^2}{\partial b} = -(y - f(x))\frac{\partial (f(x))}{\partial b} = \frac{1}{2}(f(x) - y)\frac{\partial (1 + tanh(\frac{wx+b}{2}))}{\partial b}$$

$$= \frac{1}{2}(f(x) - y)\frac{\partial (tanh(\frac{wx+b}{2}))}{\partial b} = \frac{1}{2}(f(x) - y)\frac{\partial \left(\frac{e^{\frac{wx+b}{2}} - e^{-\frac{wx+b}{2}}}{e^{\frac{wx+b}{2}} + e^{-\frac{wx+b}{2}}}\right)}{\partial b} = \frac{1}{2}(f(x) - y)\frac{\partial (\frac{e^{wx+b}-1}{e^{wx+b}+1})}{\partial b}$$

$$= \frac{f(x) - y}{2}\left[\frac{(e^{wx+b})(e^{wx+b} + 1) - (e^{wx+b})(e^{wx+b} - 1)}{(e^{wx+b} + 1)^2}\right] = (f(x) - y)\left[\frac{e^{wx+b}}{(e^{wx+b} + 1)^2}\right] \quad (2)$$

(b) $E = g(x, y, z) = \sigma(c(ax + by) + dz) = \frac{1}{1 + e^{-(c(ax+by)+dz)}} = \frac{(e^{c(ax+by)+dz})}{1 + e^{(c(ax+by)+dz)}}$

$$\implies \frac{\partial E}{\partial a} = \frac{\partial}{\partial a}\frac{e^{(c(ax+by)+dz)}}{[1 + e^{(c(ax+by)+dz)}]}$$

$$\frac{(e^{(c(ax+by)+dz)}(cx))(e^{(c(ax+by)+dz)} + 1 - e^{(c(ax+by)+dz)})}{[1 + e^{(c(ax+by)+dz)}]^2} = \frac{e^{(c(ax+by)+dz)}(cx)}{[1 + e^{(c(ax+by)+dz)}]^2} \quad (3)$$

Similarly,

$$\frac{\partial E}{\partial b} = \frac{e^{(c(ax+by)+dz)}(cy)}{[1 + e^{(c(ax+by)+dz)}]^2} \quad \text{and} \quad \frac{\partial E}{\partial d} = \frac{e^{(c(ax+by)+dz)}(d)}{[1 + e^{(c(ax+by)+dz)}]^2} \quad (4)$$

Solution 2 - Erroneous Estimates

(a) Given $f(x) = x^2 - 2x + 1$

$$\frac{df(x)}{dx} = \lim_{h\to 0}\frac{f(x + h) - f(x)}{h} = \lim_{h\to 0}\frac{(x + h)^2 - 2(x + h) + 1 - x^2 + 2x - 1}{h}$$

$$\lim_{h\to 0}\frac{h(2x-2+h)}{h} = \lim_{h\to 0}(2x-2+h) = 2x-2 \tag{5}$$

(b) Use approximation $f(x+h) = f(x) + h\dfrac{df(x)}{dx}$

$$\implies f(1+0.01) = f(1) + 0.01f'(1) = (1^2 - 2(1) + 1) + 0.01(2(1) - 2) = 0 + 0.01(0) = 0 \tag{6}$$

$$\implies f(1+0.5) = f(1) + 0.5f'(1) = (1^2 - 2(1) + 1) + 0.5(2(1) - 2) = 0 + 0.5(0) = 0 \tag{7}$$

(c) Let $\hat{f}(x)$ denote the approximation of $f(x)$. Then, the error in computing $f(0.01)$ is $f(0.01) - \hat{f}(0.01) = 0.0001$, and for $f(0.5)$, it is $f(0.5) - \hat{f}(0.5) = 0.25$.

(d) We notice discrepancy from the actual value because the approximation is defined for very small value of h ($\lim h \to 0$). This is not the case with $h = 0.01$ or $h = 0.5$. So, we need the complete Taylor series expansion.

Since $x = 1$ is a point of local minima for the function $f(x)$, the value of error increases as we move away from 1.

(e) To get a better estimate, we can use Taylor series expansion upto a higher order (second order in this case).

$$f(1.01) = f(1) + (0.01)[(2)(1) - 2] + (0.01)^2\left(\frac{2}{2!}\right) = 0.0001 \tag{8}$$

$$f(1.5) = f(1) + (0.5)[(2)(1) - 2] + (0.5)^2\left(\frac{2}{2!}\right) = 0.25 \tag{9}$$

Solution 3 - Differentiation w.r.t. Vectors and matrices

(a) To find the following gradients

$$\nabla_\times u^T x = \frac{\partial}{\partial x}(u^T x) \iff \frac{\partial}{\partial x_i}\left(\sum_{i=1}^{n} u_i x_i\right) = u_i \implies (u_1, u_2, ..., u_n) = u^T \tag{10}$$

$$\nabla_\times x^T x = \frac{\partial}{\partial x}(x^T x) \iff \frac{\partial}{\partial x_i}\left(\sum_{i=1}^{n} x_i x_i\right) = 2x_i \implies 2x^T \tag{11}$$

$$\nabla_\times x^T A x = \frac{\partial}{\partial x}(x^T A x) = \frac{\partial(x^T(Ax))}{\partial x} + \frac{\partial((x^T A)x)}{\partial x} = (Ax)^T + x^T A = x^T(A^T + A) \tag{12}$$

$$\nabla_A x^T A x = \frac{\partial}{\partial a_{ij}} \left(\begin{bmatrix} x_1 & x_2 & ... & x_n \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ & ... & & \\ a_{n1} & a_{n2} & ... & a_{nn} \end{bmatrix}^T \begin{bmatrix} x_1 \\ x_2 \\ ... \\ x_n \end{bmatrix} \right)$$

$$= \frac{\partial}{\partial a_{ij}} \left(\begin{bmatrix} a_{11}x_1 + a_{21}x_2 + ... + a_{n1}x_n \\ a_{12}x_1 + a_{22}x_2 + ... + a_{n2}x_n \\ ... \\ a_{1n}x_1 + a_{2n}x_2 + ... + a_{nn}x_n \end{bmatrix}^T \begin{bmatrix} x_1 \\ x_2 \\ ... \\ x_n \end{bmatrix} \right) = \begin{bmatrix} x_1^2 & x_1x_2 & ... & x_1x_n \\ x_1x_2 & x_2^2 & ... & x_2x_n \\ & ... & & \\ x_1x_n & x_1x_n & ... & x_n^2 \end{bmatrix} = xx^T \tag{13}$$

$$\nabla^2_{\times} x^T A x = \frac{\partial^2}{\partial x^2} x^T (Ax) = \frac{\partial}{\partial x} x^T (A^T + A) = A^T + A \tag{14}$$

(b) Let $Y = Xw + \epsilon$, where ϵ is the error residual. We define mean-squared error (MSE) as

$$\epsilon^T \epsilon = (Y - Xw)^T (Y - Xw) = Y^T Y - 2wX^T Y + wX^T Xw \quad [\because (wX^T Y)^T = Y^T Xw]$$

$$\therefore \frac{\partial}{\partial w}(\epsilon^T \epsilon) = -2X^T Y + 2X^T Xw = 0 \implies X^T Y = X^T Xw \implies w = (X^T X)^{-1} X^T Y \tag{15}$$

(c) We can write $\nabla_T f$ in terms of n 2-dimensional arrays $(A_1, A_2, A_3, ..., A_n)$ as

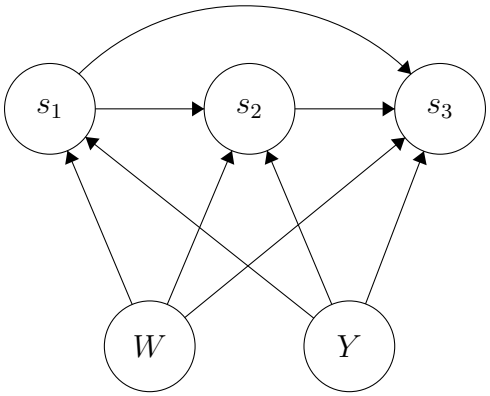
$$\nabla_T f = \begin{bmatrix} \nabla_T A_1 & \nabla_T A_2 & \nabla_T A_3 & ... & \nabla_T A_n \end{bmatrix} \tag{16}$$

Solution 4 - Ordered Derivatives

(a) Given $y_4(y_3, y_2), y_3(y_2, y_1), y_2(y_1)$

$$\frac{\partial y_4}{\partial y_1} = \frac{\partial y_4}{\partial y_3}\frac{\partial y_3}{\partial y_1} + \frac{\partial y_4}{\partial y_2}\frac{\partial y_2}{\partial y_1} = \frac{\partial y_4}{\partial y_3}\frac{\partial y_3}{\partial y_2}\frac{\partial y_2}{\partial y_1} + \frac{\partial y_4}{\partial y_3}\frac{\partial y_3}{\partial y_1} + \frac{\partial y_4}{\partial y_2}\frac{\partial y_2}{\partial y_1} = \frac{\partial y_4}{\partial y_3}\left[\frac{\partial y_3}{\partial y_2}\frac{\partial y_2}{\partial y_1} + \frac{\partial y_3}{\partial y_1}\right] + \frac{\partial y_4}{\partial y_2}\frac{\partial y_2}{\partial y_1}$$

(b) The dependency graph for the variables s_3, s_2, s_1, W, Y is as follows.



(c) We can pbtain the expressions using the chain rule of differentiation as per the given hierarchy in the dependency graph.

$$\frac{\partial s_3}{\partial W} = s_3(1-s_3)(s_2 + s_2(1-s_2)(s_1 + s_0 + Ws_{-1}) + Y(s_0 + Ws_{-1}))$$

$$\frac{\partial s_3}{\partial Y} = s_3(1-s_3)(Ws_2(1-s_2)(Ws_1(1-s_1)s_{-1} + s_0) + Y(s_1(1-s_1)s_{-1} + s_0)) + s_1)$$

$$\frac{\partial s_3}{\partial U} = s_3(1-s_3)x_3$$

Solution 5 - Baby Steps

(a) We can use an approach like gradient descent and repeat it until convergence.

$$x_t := x_t - \alpha \frac{d}{dx} f(x) \quad \text{where } t = 0, ..., n \text{ and } \alpha \text{ is the learning rate}$$

(b) For multivariable case, we can use the following approach until convergence.

$$x_t := x_t - \alpha \frac{d}{dx} f(x, y) \quad \text{and} \quad y_t := y_t - \alpha \frac{d}{dy} f(x, y)$$

(c) This method can only be used to find a local minima and does not guarantee the global minima. It may fail if the function is non-convex. The above approach can also lead to a different minima depending upon the initial parameter x_0. If we choose a small value of α, it may take very large time to converge. Or it can lead to a steep local minima instead of a gradual global minima.

(d) The above approach will always work for convex functions because there is only a single minima (global), so the problem of leading to a local minima will not be applicable.

(e) The number of steps depend upon the learning rate. If α is small, then the number of steps will be more, while a large α will require less steps. However, in case α is large, we may overshoot the minima and iterations may not converge.

(f) The number of steps can be improved by improving the learning rate. We can set α to be a function of iterations, such that larger α in the beginning (for less steps) and it gets smaller as the iterations progress (for convergence). Other methods, such as Nestrov momentum, can also be used to reduce the number of steps.

Solution 6 - Constrained Optimization

(a) We can understand this by analysing the nature of the slope. At the point of minima slope changes from negative to positive. During this transition, slope will become zero. And slope will be zero with respect to every variable. The gradient of a multivariable function at a minima will be the zero vector, which corresponds to the graph having a flat tangent plane. So, concept of considering single variable at a time for partial derivative indicates the possibility of a minima.

(b) To make sure that the above method works, we need to represent $g(x, y) = c$ in terms of $x = u(y)$ or $y = v(x)$ and replace x or y in f respectively. If this is not possible, then we need to use the Lagrange multiplier method for constraint optimization.

(c) The component of ∇g along the feasible curve is perpendicular. So, the component is zero.

(d) The component of ∇f (gradient) at minima of the curve is perpendicular. So, it is zero.

(e) As seen above, ∇g and ∇f will be in the same direction perpendicular to the feasible curve, at the minima. Hence, $\nabla f \propto \nabla g$ or $\nabla f = \lambda \nabla g$. Here, λ is the Lagrange multiplier.

We can also see that the gradient of a function is perpendicular to the contour lines. The contour lines of f and g are parallel if and only if the gradients of functions are parallel (i.e. $\nabla f = \lambda \nabla g$). Thus we want points (x, y) where $g(x, y) = 0$.

(f) Let $\mathcal{L}(x, y, \lambda) = x^a y^b z^c - \lambda(x + y + z - 1)$. The contour lines of $x^a y^b z^c$ and $(x + y + z - 1)$ are parallel if and only if the gradients of the functions are parallel. So, want the points (x, y, z), i.e. $\nabla_{x,y,z,\lambda} \mathcal{L}(x, y, \lambda) = 0$.

$$\implies \nabla_{x,y,z}(x^a y^b z^c) = \lambda \nabla_{x,y,z}(x+y+z-1) \quad \text{and} \quad \nabla_\lambda \mathcal{L}(x, y, \lambda) = 0 \quad \text{i.e.} \quad x+y+z-1 = 0 \tag{17}$$

So, we have to solve the following simultaneous equations to find (x, y, z).

$$a x^{a-1} y^b z^c = b x^a y^{b-1} z^c = c x^a y^b z^{c-1} = \lambda \quad (\text{from } \nabla_{x,y,z} \mathcal{L}(x, y, \lambda) = 0) \tag{18}$$

$$\implies x = \frac{a}{a+b+c} \quad , \quad y = \frac{b}{a+b+c} \quad , \quad z = \frac{c}{a+b+c}$$

$$\therefore \max_{x,y,z} x^a y^b z^c = \frac{a^a b^b c^c}{(a+b+c)^{a+b+c}} \tag{19}$$

Solution 7 - Billions of Balloons

(a) First, we define the error and the accuracy based on our estimates $\hat{k}_1, \hat{k}_2, \hat{k}_3$.

$$\text{Error } (\epsilon) = \frac{|10^6\hat{k}_1 - k_1| + |10^6\hat{k}_2 - k_2| + |10^6\hat{k}_3 - k_3|}{10^9} \quad \text{and} \quad \text{Accuracy} = 1 - \epsilon$$

Now, $P(k_i) = \dfrac{k_i}{\sum_i k_i}$. Hence, we can quantify the deviation from actual value by calculating the probability difference $= |P(\hat{k}_1) - P(k_1)| + |P(\hat{k}_2) - P(k_2)| + |P(\hat{k}_3) - P(k_3)|$.

(b) $\mathbf{q}$ is easy to use for optimization because of its differentiability. Hence, it is good for 1-N classification. q_i is the Softmax classifier and it can be used to determine any N-class probability function over the feature space (i.e. the Universal Approximation Theorem). But, $\mathbf{q}$ might give a higher estimate for the more frequent quantity and a lesser estimate for the less frequent.

Solution 8 - Expectation

In general, the expectation of a CDF (here, $F_X(x)$) is defined as follows.

$$\mathbb{E}_X[F(X)] = \int_{-\infty}^{\infty} F(x)p(x)\,dx \quad \text{where} \quad p(x) = \frac{d}{dx}F_X(x)$$

$$\implies \mathbb{E}_X[F(X)] = \int_{-\infty}^{\infty} F(x)\,d(F_X(x)) = \int_0^1 F(x)\,d(F_X(x)) \quad [\because F_X(x)\ (\text{CDF}) \in (0,1)]$$

$$\implies \mathbb{E}_X[F(X)] = \left[\frac{1}{2}F_X^2(x)\right]_{F_X(x)=0}^{F_X(x)=1} = \frac{1}{2} \tag{20}$$

Solution 9 - Intuitive Urns

(a) Bob seems to have stronger evidence since he had performed the experiment more number of times and has a larger sample. So, his error probability is less compared to Alice.

(b) We can use the Bayes' Theorem here.

$$P\left(\frac{Blue}{Alice}\right) = \left[P\left(\frac{Alice}{Blue}\right) \times P(Blue)\right] \;/\; \left[P\left(\frac{Alice}{Blue}\right) \times P(Blue) + P\left(\frac{Alice}{Red}\right) \times P(Red)\right]$$

$$= \left[\left(\frac{3}{5}\right)^6 \frac{1}{2}\right] \;/\; \left[\left(\frac{3}{5}\right)^6 \frac{1}{2} + \left(\frac{2}{5}\right)^6 \frac{1}{2}\right] = \frac{(3)^6}{(3)^6 + (2)^6} = \frac{729}{793} \tag{21}$$

(c) We can again use the Bayes' Theorem here.

$$P\left(\frac{Blue}{Bob}\right) = \left[P\left(\frac{Bob}{Blue}\right) \times P(Blue)\right] \;/\; \left[P\left(\frac{Bob}{Blue}\right) \times P(Blue) + P\left(\frac{Bob}{Red}\right) \times P(Red)\right]$$

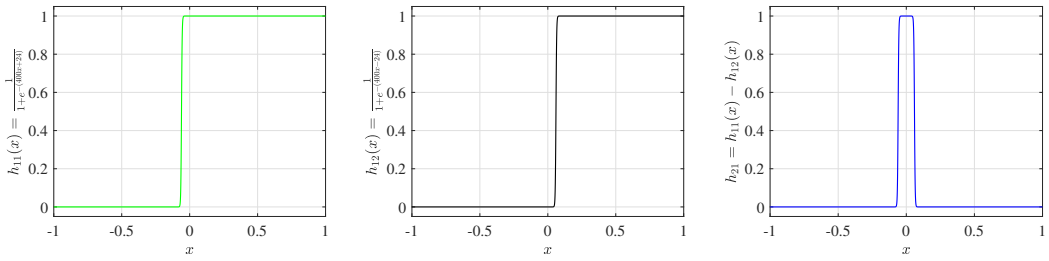
$$= \left[{}^{600}C_{297}\left(\frac{3}{5}\right)^{303}\left(\frac{2}{5}\right)^{297}\frac{1}{2}\right] \;/\; \left[{}^{600}C_{297}\left(\frac{3}{5}\right)^{303}\left(\frac{2}{5}\right)^{297}\frac{1}{2} + {}^{600}C_{303}\left(\frac{2}{5}\right)^{303}\left(\frac{3}{5}\right)^{297}\frac{1}{2}\right]$$

$$= \frac{(3)^6}{(3)^6 + (2)^6} = \frac{729}{793} \tag{22}$$

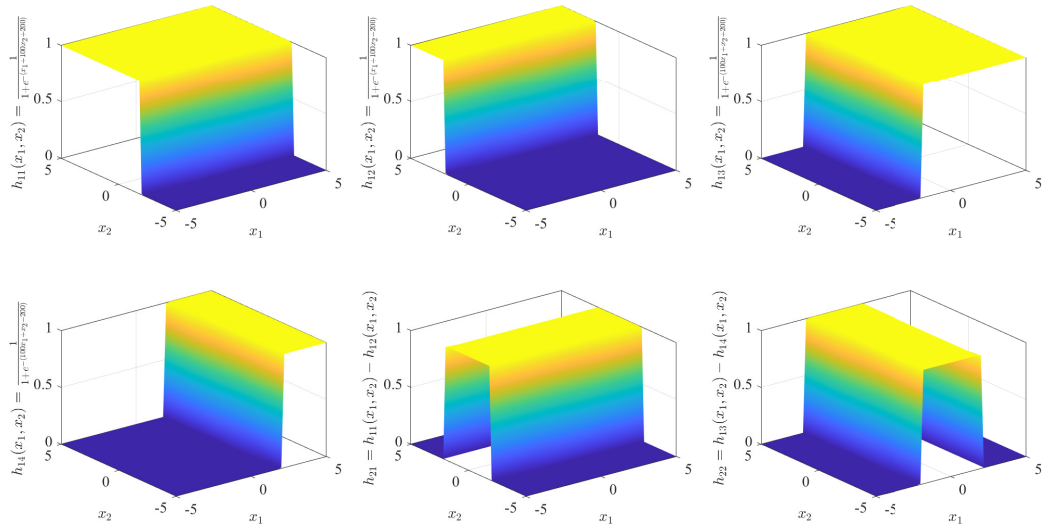
(d) As seen above, Alice and Bob both have the same probability of being correct. Therefore, both have strong evidence for claiming that the removed ball was blue. Hence, our initial intuition that Bob having a lesser error probability, was wrong.

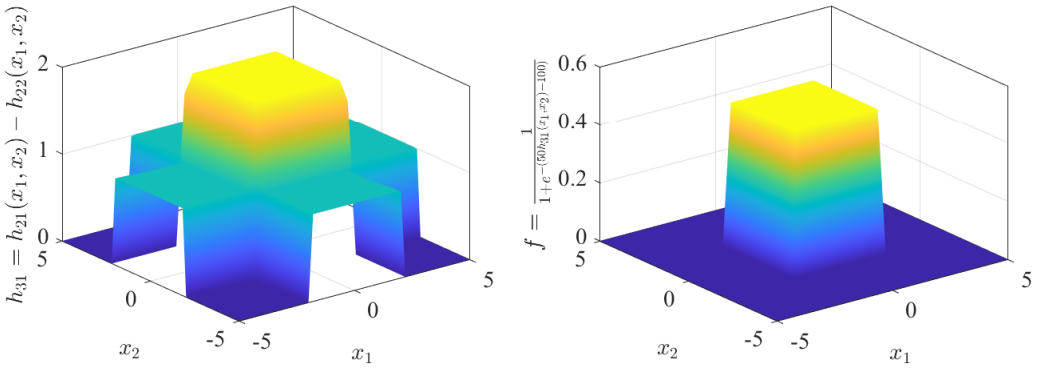
Solution 10 - Plotting Functions for Great Good

(a) Plots of $h_{11}(x), h_{12}(x)$ and $h_{21}(x)$ for $x \in (-1, 1)$ are shown below.



(b) Plots of $h_{11}(x_1, x_2), h_{12}(x_1, x_2), h_{13}(x_1, x_2), h_{14}(x_1, x_2), h_{21}(x_1, x_2), h_{22}(x_1, x_2), h_{31}(x_1, x_2)$ and $f(x_1, x_2)$ for $x_1 \in (-5, 5)$ and $x_2 \in (-5, 5)$ are shown below.





References

[1] Mathematics for Machine Learning | Companion webpage to the book.
https://mml-book.com/

[2] Single Variable Calculus | Mathematics | MIT OpenCourseWare.
https://ocw.mit.edu/courses/18-01-single-variable-calculus-fall-2006/

[3] Multivariable Calculus | Mathematics | MIT OpenCourseWare.
https://ocw.mit.edu/courses/18-02-multivariable-calculus-fall-2007/