CSE253: Homework #1

Due on 11 59 pm, Jan 20, 2020 at 11:59pm

Prof. Gary W. Cottrell

Sai Akhil Suggu, A53284020

Problem 1

Consider the follwing identity involving the transformation from Cartesian to polar coordinates

$$\prod_{i=1}^{d}\int_{-\infty}^{\infty}\exp(-x^2)dx = S_d\int_{0}^{\infty}\exp(-r^2)r^{d-1}dr \tag{1}$$

where S_d is the surface area of the unit sphere in d dimentions. By making use of

$$\int_{-\infty}^{\infty}\exp(\frac{-\lambda x^2}{2})dx = (\frac{2\pi}{\lambda})^{\frac{1}{2}} \tag{2}$$

Show that

$$S_d = \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})}$$

where $\Gamma(x)$ is the Gamma function defined by

$$\Gamma(x) = \int_{0}^{\infty} u^{x-1}exp(-u)du \tag{3}$$

Solution
Using (2), when $\lambda = 2$, we have

$$\int_{-\infty}^{\infty}\exp(-x^2)dx = (\pi)^{\frac{1}{2}} \tag{4}$$

which imples L.H.S of (1) is

$$\prod_{i=1}^{d}\int_{-\infty}^{\infty}\exp(-x^2)dx = (\pi)^{\frac{d}{2}} \tag{5}$$

Considering the integral in R.H.S of (1),

$$\begin{aligned}\int_{0}^{\infty}\exp(-r^2)r^{d-1}dr &= \int_{0}^{\infty}\exp(-u)u^{\frac{d-1}{2}}\cdot\frac{du}{2\sqrt{u}} && \text{(with substitution } u = r^2\text{)}\\ &= \frac{\int_{0}^{\infty}\exp(-u)u^{\frac{d}{2}-1}du}{2}\\ &= \frac{\Gamma(\frac{d}{2})}{2} && \text{using (3)}\end{aligned}$$

Substituting above derived values in both sides of (1),

$$(\pi)^{\frac{d}{2}} = S_d\frac{\Gamma(\frac{d}{2})}{2}$$

Thus,

$$S_d = \frac{2(\pi)^{\frac{d}{2}}}{\Gamma(\frac{d}{2})}$$

Hence prooved the required equation

Sanity check :
When d = 2, circumference of circle $S_d = \frac{2\pi^{\frac{2}{2}}}{\Gamma(1)} = 2\pi$, consistent with our knowledge
When d = 3, surface area of sphere $S_d = \frac{2\pi^{\frac{3}{2}}}{\Gamma(\frac{3}{2})} = 4\pi$, consistent with our knowledge

Problem 2

Show that the volume of a hypersphere of radius a in d-dimentions is given by

$$V_d = \frac{S_d a^d}{d} \tag{6}$$

Hence show that the ratio of the volume of a hypersphere of radius a to the volume of hupercude of side 2a (i.e, circumscribed hypercube) is given by

$$\frac{\text{Volume of sphere}}{\text{Volume of cube}} = \frac{\pi^{\frac{d}{2}}}{d2^{d-1}\Gamma(\frac{d}{2})} \tag{7}$$

using Stirlings approximation

$$\Gamma(x+1) = (2\pi)^{\frac{1}{2}} \exp(-x) x^{\frac{x+1}{2}} \tag{8}$$

which is valid when x is large, show that as $d \to \infty$, the ratio in (7) goes to zero. Similarly, show that the ratio of distance from centre of the hypercube to one of the corners, divided by the perpendicular distance to one of the faces is $\sqrt{d}$, and therefore goes to ∞ as $d \to \infty$. These results show that, in a high dimentional space, most of the volume of a cube is concentrated in the large number of corners which thmeselves become very long spikes

Solution

Volume of a hypersphere can be found by integrating the the extra volume (marginal volume) we get when radius increases from r to r+dr

$$\begin{aligned}\text{Volume of hypersphere} &= \int_0^a \text{marginal volume (r} \to \text{r+dr)} \\ &= \int_0^a S_d(r) dr\end{aligned}$$

From problem 1, we have

$$\begin{aligned}S_d(a) &= \frac{2\pi^{d/2} a^{d-1}}{\Gamma(\frac{d}{2})} && \text{substituting r = ak, we get extra multiplier for } a^{d-1} \text{ for radius a} \\ &= S_d a^{d-1}\end{aligned}$$

Thus,

$$\begin{aligned}\text{Volume of hypersphere} &= \int_0^a S_d r^{d-1} dr \\ &= \frac{S_d a^d}{d} && \text{Q.E.D for (6)}\end{aligned}$$

Volume od hypercube (of length 2a) = $(2a)^d$
which implies

$$\begin{aligned}\frac{\text{Volume of sphere}}{\text{Volume of cube}} &= \frac{S_d a^d}{d(2a)^d} \qquad (9) \\ &= \frac{\pi^{\frac{d}{2}}}{d2^{d-1}\Gamma(\frac{d}{2})} && \text{(substituing values calculated earlier)} \qquad (10)\end{aligned}$$

Problem 2 continued on next page. . .

Thus, eqn (7) is prooved.

As $d \rightarrow \infty$, denominator in above equation goes to zero. From Stirling approximation, $(\frac{2}{\pi})^n\Gamma(n/2) \approx \exp(-n/2)(\frac{n-2}{2})^{n/2}$. When $n > e$, multiplier term is great than one. Thus, total product goes infinity. Thus, R.H.S of eqn (7) goes to zero.

In a hypercube of length 2 with centred at (0,0,0,0...0) $\in \mathbb{R}^d$,

From symmetry aorund center, all corners are at equidistance from the centre. Taking the corner (1,1,1,.....1) $\in \mathbb{R}^d$, distance from centre is $\sqrt{d}$ perpendicular from centre to face lies on coordinate axis. Thus, W.L.O.G coordinate is (1,0,0,0,0 ...0) who se distance from the centre is 1.

Hence ratio of these two distance is $= \sqrt{d}$ which goes to ∞ as $d \rightarrow \infty$
Therefore, as number of dimensions increase, most of the volume of a cube is concentrated in the large number of corners which thmeselves become very long spikes

Problem 3

from Bishop Neural Networks for Pattern recognition 1.3

Solution
Using (6), we have Volume of Sprere of radius a,
$V_d(a) = \frac{S_d a^d}{d}$
$V_d(a-\epsilon) = \frac{S_d (a-\epsilon)^d}{d}$
Thus, Required fraction of volume,

$$f = \frac{V_d(a) - V_d(a-\epsilon)}{V_d(a)}$$
$$= 1 - (\frac{a-\epsilon}{a})^d$$
$$= 1 - (1 - \frac{\epsilon}{a})^d$$

Now, for any value of $\epsilon > 0$, $0 < \epsilon \leq a$, $(1 - \frac{\epsilon}{a}) < 1$.
Thus as d $\rightarrow \infty, (1 - \frac{\epsilon}{a}) \rightarrow 0$.

Thus, the fraction $f \rightarrow 1$.

$$\lim_{d \rightarrow \infty} f = 1$$

,
For $\frac{\epsilon}{a} = 0.01$, Evaluating f numerically,
for d = 2, f = 0.0199
for d =10, f = 0.0297
for d =1000, f = 0.9999

for fraction of volume of the sphere indie radius = $\frac{a}{2}$, we can consider $\epsilon = \frac{a}{2}$, and for answer we need 1 - f
Thus, For $\frac{\epsilon}{a} = 0.5$, Evaluating f numerically,
for d = 2, f = 0.75, 1-f = 0.25
for d =10, f = 0.875, 1-f = 0.125
for d =1000, $f \approx 1, 1 - f = 9.33 \cdot 10^{-302}$

Above can be caluated much easily using $f_{\frac{a}{2}} = (0.5)^d$
We observe that in higher dimentions, most of the volume is near outer layer. Hence, when we sample, almost all points are concentrated near in a thin cell close to the surface

Problem 4

from Bishop Neural Networks for Pattern recognition 1.4

Solution

Given probability density function,

$$p(x) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp(-\frac{\|x\|^2}{2\sigma^2})$$

Converting to polar coordinates, we have,

$$p(r) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp(-\frac{r^2}{2\sigma^2})$$

Now, probability mass inside a thin shell of radius ϵ, is give by,

$$\rho(r)\epsilon = p(r)(V_d(r+\epsilon) - V_d(r)) = p(r)S_d(r)\epsilon$$

as seen in problem 2, $S_d(r) = S_d r^{d-1}$

$$\rho(r) = p(r)S_d(r) = \frac{S_d r^{d-1}}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp(-\frac{r^2}{2\sigma^2})$$

for maximum value of $\rho(r)$,

$$\begin{aligned}
\frac{d\rho(r)}{dr} &= 0 \\
\frac{d(r^{d-1}\exp(-\frac{r^2}{2\sigma^2}))}{r} &= 0 \\
(d-1)r^{d-2}\exp(-\frac{r^2}{2\sigma^2}) + r^{d-1}\exp(-\frac{r^2}{2\sigma^2}) - \frac{r}{\sigma^2} &= 0 \\
(d-1) - \frac{r^2}{\sigma^2} &= 0 \\
\dot{r} &= \sigma\sqrt{d-1}
\end{aligned}$$

Observe that,

$$\frac{d^2\rho(r)}{dr^2} = \frac{-2k}{\sigma^2} < 0$$

Thus, slope is always decreaing. hence we have only ane extrema and function is maximum at that point

For large values of d, probability mass is maximum at

$$\dot{r} \approx \sigma\sqrt{d}$$

Approximating the thin cell :
Considering the ratio,

$$\begin{aligned}
\frac{\rho(r+\epsilon)}{\rho(r)} &= (\frac{r+\epsilon}{r})^{d-1} exp(\frac{r^2}{2\sigma^2} - \frac{(r+\epsilon)^2}{2\sigma^2}) \\
&= (1+\frac{\epsilon}{r})^{d-1} exp(-\frac{2r\epsilon+\epsilon^2}{2\sigma^2}) \\
&= \exp((d-1)\ln(1+\frac{\epsilon}{r}) - \frac{d\epsilon}{r} - \frac{d\epsilon^2}{2r^2})
\end{aligned}$$

Problem 4 continued on next page. . .

When $\epsilon \ll r \implies \frac{\epsilon}{r} \approx 0 \implies \ln(1+\frac{\epsilon}{r}) = \frac{\epsilon}{r} - \frac{\epsilon^2}{r^2}$
Thus, when d is large and as $ln(1+x) - x \approx \frac{-x^2}{2}$,
$(d-1)\ln(1+\frac{\epsilon}{r}) - \frac{d\epsilon}{r} \approx -d\frac{\epsilon^2}{2r^2} = \frac{\epsilon^2}{\sigma^2}$
Which implies, $\frac{\rho(r+\epsilon)}{\rho(r)} = \exp(-\frac{\epsilon^2}{\sigma^2})$

$$\rho(r+\epsilon) = \rho(r)\exp(-\frac{\epsilon^2}{\sigma^2})$$

which says that $\rho(r)$ decays exponentially away from its maximum

Also note that probability denisty is maximum at r =0, but probability mass is maximum at r = $\dot{r}$. That is bulk of probability mass is located in a different part of space from the region of high probability denisty

Problem 5

Logistic Regression

Logistic regression is a binary classification method. Intuitively, logistic regression can be conceptualized as a single neuron reading in a d-dimensional input vector $x \in \mathbb{R}^d$ and producing an output y between 0 and 1 that is the system's estimate of the conditional probability that the input is in some target category. The "neuron" is parameterized by a weight vector $w \in \mathbb{R}^{d+1}$, where w_0 represents the bias term (a weight from a unit that has a constant value of 1).

Consider the following model parametrized by w:

$$y = P(C_1 \mid x) = \frac{1}{1 + \exp(-w^\top x)} = g(w^\top x) \tag{11}$$

$$P(C_0 \mid x) = 1 - P(C_1 \mid x) = 1 - y \tag{12}$$

where we assume that x has been augmented by a leading 1 to represent the bias input. With the model so defined, we now define the Cross-Entropy cost function, equation (13), the quantity we want to minimize over our training examples:

$$\mathbb{E}[w] = -\sum_{n=1}^{N} \{t^n \ln(y^n) + (1 - t^n) \ln(1 - y^n)\} \tag{13}$$

Here, $t^n \in \{0, 1\}$ is the label or teaching signal for example n ($t^n = 1$ represents $x^n \in C_1$). We minimize this cost function via gradient descent.

To do so, we need to derive the gradient of the cost function with respect to the parameters w_j. Assuming we use the logistic activation function g as in equation (11), prove that this gradient is:

$$-\frac{\partial E}{\partial w_j} = \sum_{n=1}^{N} (t^n - y^n) x_j^n \tag{14}$$

Solution

To calculate the required derivative, let's first calculate few other required derivates

$$\frac{\partial y}{\partial w_j} = \frac{\partial g(z)}{\partial z} x_j \qquad \text{taking } z = w^\top x$$

$$\begin{aligned}
\frac{\partial y}{\partial w_j} &= \frac{\partial g(z)}{\partial z} x_j && \text{taking } z = w^\top x \\
\frac{\partial g(z)}{\partial z} &= \frac{\exp(-z)}{(1 + \exp(-z))^2} && \text{testing} \\
&= \frac{1}{1 + \exp(-z)} \cdot \frac{exp(-z)}{1 + \exp(-z)} \\
&= g(z)(1 - g(z))
\end{aligned}$$

$$\begin{aligned}
\frac{\partial y}{\partial w_j} &= \frac{\partial g(z)}{\partial z}x_j && \text{taking } z = w^\top x \\
\frac{\partial g(z)}{\partial z} &= \frac{\exp(-z)}{(1+\exp(-z))^2} && \text{testing} \\
&= \frac{1}{1+\exp(-z)} \cdot \frac{exp(-z)}{1+\exp(-z)} \\
&= g(z)(1-g(z))
\end{aligned}$$

which imples,

$$\frac{\partial y^n}{\partial w_j} = y^n(1-y^n)$$

finally,

$$\begin{aligned}
-\frac{\partial E}{\partial w_j} &= \sum_{n=1}^{N}\{\frac{t^n}{y^n}\frac{\partial y^n}{\partial w_j} + \frac{1-t^n}{1-y^n}\frac{-\partial y^n}{\partial w_j}\} \\
&= \sum_{n=1}^{N}\{\frac{t^n}{y^n}y^n(1-y^n) - \frac{1-t^n}{1-y^n}y^n(1-y^n)\}x_j \\
&= \sum_{n=1}^{N}\{t^n(1-y^n) - (1-t^n)y^n\}x_j \\
&= \sum_{n=1}^{N}\{t^n - y^n\}x_j && \text{Q.E.D}
\end{aligned}$$