Deep Learning/Deep Learning for Vision

Vineeth N Balasubramanian
Indian Institute of Technology, Hyderabad

ASSIGNMENT 0

This is a preparatory assignment for the course on "Deep Learning"/"Deep Learning for Vision". Below are questions on basics/pre-requisites for this course. If you can answer these, you are ready for the course!

1 Functions and Derivatives

1.1 Simple derivatives

Question: Find the derivative of $(\sin x + e^{2x} + \sqrt{x})$.
Solution:

$$\begin{aligned}\frac{d}{dx}(\sin x + e^{2x} + \sqrt{x}) &= \frac{d}{dx}\sin x + \frac{d}{dx}e^{2x} + \frac{d}{dx}\sqrt{x} \\ &= \cos x + e^{2x}\frac{d}{dx}(2x) + \frac{1}{2}x^{\frac{1}{2}-1} \\ &= \cos x + 2e^{2x} + \frac{1}{2}x^{-\frac{1}{2}} \\ &= \cos x + 2e^{2x} + \frac{1}{2\sqrt{x}}\end{aligned}$$

1.2 Activation function

Question: Activation functions are used by neural networks to learn non-linear decision boundaries. Popular activation functions used in neural networks are sigmoid, tanh and ReLU. Find the derivative of $\tanh(x)$. Recall that $\tanh(x)$ is defined as:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

Solution:

$$\begin{aligned}
\frac{d}{dx}(\tanh(x)) &= \frac{d}{dx}(\frac{e^x - e^{-x}}{e^x + e^{-x}}) \\
&= \frac{(e^x + e^{-x})(e^x + e^{-x}) - (e^x - e^{-x})(e^x - e^{-x})}{(e^x + e^{-x})^2} \\
&= \frac{(e^x + e^{-x})^2 - (e^x - e^{-x})^2}{(e^x + e^{-x})^2} \\
&= 1 - \frac{(e^x - e^{-x})^2}{(e^x + e^{-x})^2} \\
&= 1 - (\frac{e^x - e^{-x}}{e^x + e^{-x}})^2 \\
&= 1 - (\tanh(x))^2
\end{aligned}$$

1.3 Chain rule

Question: Apply chain rule to find the derivative of $\sin \frac{\sqrt{e^x+a}}{2}$. where a is constant.
Solution:

$$\begin{aligned}
\frac{d}{dx} \sin \frac{\sqrt{e^x + a}}{2} &= \cos \frac{\sqrt{e^x + a}}{2} \frac{d}{dx}(\frac{\sqrt{e^x + a}}{2}) \\
&= \cos \frac{\sqrt{e^x + a}}{2} \frac{1}{2} \frac{d}{dx}(\sqrt{e^x + a}) \\
&= \cos \frac{\sqrt{e^x + a}}{2} \frac{1}{2} \frac{1}{2\sqrt{e^x + a}} \frac{d}{dx}(e^x + a) \\
&= \cos \frac{\sqrt{e^x + a}}{2} \frac{1}{2} \frac{1}{2\sqrt{e^x + a}}(e^x + 0) \\
&= \cos \frac{\sqrt{e^x + a}}{2} \frac{e^x}{4\sqrt{e^x + a}}
\end{aligned}$$

2 Probability and Statistics

2.1 Bayes Theorem

Question: Using Bayes theorem, answer the following question. A laboratory blood test is 95% effective in detecting a certain disease when it is, in fact, present. However, the test also yields a "false positive" result for 1% of healthy persons tested (i.e., if a healthy person is tested, then with probability 0.01, the test result will imply that he or she has the disease). If 0.5% of the population actually has the disease, what is the probability that a person has the disease given that the test result is positive?

Solution:

Let E represent the event that the lab test outputs the presence of disease (tests positive), E^c represent the complementary event that lab test outputs the absence of disease (tests negative), D represent the event that the individual is actually diseased, and D^c represent the event that the individual is actually not diseased. It is given that the test is 95% effective if the person is actually diseased,i.e. $P(E|D) = 0.95$. Similarly, the test yields a false positive result for 1% of healthy persons tested, i.e. $P(E|D^c) = 0.01$. It is also given that in the entire population, 0.5% of the people are actually diseased, i.e. $P(D) = 0.005$. Based on this information we are asked to find the probability that a person has the disease given that the lab test given positive test result, i.e. we need to find $P(D|E)$.

Using Bayes theorem:

$$\begin{aligned} P(D|E) &= \frac{P(E|D)P(D)}{P(E)} \\ &= \frac{P(E|D)P(D)}{P(E|D)P(D) + P(E|D^c)P(D^c)} (law\ of\ total\ probability) \\ &= \frac{0.95 * 0.005}{0.95 * 0.005 + 0.01 * 0.995} \\ &= \frac{0.00475}{0.0147} \approx 0.323 \end{aligned}$$

2.2 Conditional Dependence and Conditional Independence

Recall that two random variables X, Y are independent if $P(X, Y) = P(X)P(Y)$ or $P(X|Y) = P(X)$. Similarly we can define conditional independence as follows. Two random variables X, Y are conditionally independent given another random variable Z with $P(Z) > 0$, if $P(X, Y|Z) = P(X|Z)P(Y|Z)$ or $P(X|Y, Z) = P(X|Z)$. Two random variables may be unconditionally dependent but it is possible that they are conditionally independent and vice versa. Now, answer the following question.

Question: A box contain two coins. First coin is a fair coin ($P(H) = 0.5$). Second coin is a biased coin with $P(H) = 1$. Now, we choose a coin at random and toss it twice. Let us define the following events.

- A = First toss results in head
- B = Second toss results in head
- C = Fair coin is selected

Find $P(A), P(B), P(A \cap B), P(A|C), P(B|C), P(A \cap B|C)$ and identify the relationship between A, B, C.

Solution:

It is easy to see that, given that we have a fair coin, probability of head is $\frac{1}{2}$.

$$P(A|C) = \frac{1}{2}$$

$$P(B|C) = \frac{1}{2}$$

and

$$P(A \cap B|C) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

The above statement is true because for any given coin, consecutive tosses are independent.

We now use the law of total probability to find $P(A), P(B), P(A \cap B)$:

$$\begin{aligned} P(A) &= P(A|C)P(C) + P(A|C^c)P(C^C) \\ &= \frac{1}{2} \times \frac{1}{2} + 1 \times \frac{1}{2} \\ &= \frac{3}{4} \end{aligned}$$

Similarly,

$$\begin{aligned} P(B) &= P(B|C)P(C) + P(B|C^c)P(C^C) \\ &= \frac{1}{2} \times \frac{1}{2} + 1 \times \frac{1}{2} \\ &= \frac{3}{4} \end{aligned}$$

$$\begin{aligned} P(A \cap B) &= P(A \cap B|C)P(C) + P(A \cap B|C^c)P(C^C) \\ &= P(A|C)P(B|C)P(C) + P(A|C)P(B^c|C^c)P(C^c) \\ &= \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} + 1 \times 1 \times \frac{1}{2} \\ &= \frac{5}{8} \end{aligned}$$

From the above calculations, we can see that: $P(A \cap B|C) = \frac{1}{4} = P(A|C)P(B|C) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$ and

$$P(A \cap B) = \frac{5}{8} \neq P(A)P(B) = \frac{9}{16}$$

That is, events A, B are originally dependent but after conditioning on C, they became independent of each other.

2.3 Entropy, Cross Entropy and KL divergence

Question: Write the expression of KL divergence in terms of entropy and cross-entropy. Comment on the 'symmetry' of KL divergence (i.e. is $KL(P, Q) = KL(Q, P)$?).
Solution:

KL divergence between two probability distributions P, Q is defined as:

$$D_{KL}(P||Q) = H(P, Q) - H(P)$$

where

$$H(P, Q) = -\sum_{x \sim P} P(x) \log(Q(x))$$

is the cross entropy between P and Q and

$$H(P) = -\sum_{x \sim P} P(x) \log(P(x))$$

is the entropy of the random variable X having probability distribution P.

$$\begin{aligned} D_{KL}(P||Q) &= H(P,Q) - H(P) \\ &= -\sum_{x \sim P} P(x) \log(Q(x)) - (-\sum_{x \sim P} P(x) \log(P(x))) \\ &= -\sum_{x \sim P} P(x) \log(Q(x)) + \sum_{x \sim P} P(x) \log(P(x)) \\ &= \sum_{x \sim P} P(x) \log\left(\frac{P(x)}{Q(x)}\right) \end{aligned}$$

$$D_{KL}(P||Q) = \sum_{x \sim P} P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

Now consider $D_{KL}(Q||P)$:

$$\begin{aligned} D_{KL}(Q||P) &= H(Q,P) - H(Q) \\ &= -\sum_{x \sim Q} Q(x) \log(P(x)) - (-\sum_{x \sim Q} Q(x) \log(Q(x))) \\ &= -\sum_{x \sim Q} Q(x) \log(P(x)) + \sum_{x \sim Q} Q(x) \log(Q(x)) \\ &= \sum_{x \sim Q} Q(x) \log\left(\frac{Q(x)}{P(x)}\right) \\ &\neq D_{KL}(P||Q) \end{aligned}$$

So, KL divergence is not symmetric.

$$\boxed{D_{KL}(P||Q) \neq D_{KL}(Q||P)}$$

Question: Find the KL divergence between two probability distributions P, Q where P, Q are as follows:

X	x_1	x_2	x_3
$P(X = x_i)$	0	0	1
$Q(X = x_i)$	0.25	0.5	0.25

Solution:

Using the formula for calculating KL divergence, we have:

$$
\begin{aligned}
D_{KL}(P||Q) &= \sum_{x\sim P} P(x)\log\left(\frac{P(x)}{Q(x)}\right) \\
&= 0\log\left(\frac{0}{0.25}\right) + 0\log\left(\frac{0}{0.5}\right) + 1\log\left(\frac{1}{0.25}\right) \\
&= \log(4)
\end{aligned}
$$

2.4 Mutual Information

Question: Mutual Information of two random variables X, Y is defined as follows:

$$I(X;Y) = \sum_{x\in\mathbb{X},y\in\mathbb{Y}} P_{(X,Y)}(x,y)\log\frac{P_{(X,Y)}(x,y)}{P_X(x)P_Y(y)}$$

Prove that $I(X;Y) = H(Y) - H(Y|X)$, where $H(Y)$ is entropy of random variable Y and $H(Y|X)$ is conditional entropy of Y given X.
Solution:

$$
\begin{aligned}
I(X;Y) &= \sum_{x\in\mathbb{X},y\in\mathbb{Y}} P_{(\mathbf{X},\mathbf{Y})}(x,y)\log\frac{P_{(\mathbf{X},\mathbf{Y})}(x,y)}{P_{\mathbf{X}}(x)P_{\mathbf{Y}}(y)} \\
&= \sum_{x\in\mathbb{X},y\in\mathbb{Y}} P_{(\mathbf{X},\mathbf{Y})}(x,y)\log\frac{P_{(\mathbf{X},\mathbf{Y})}(x,y)}{P_{\mathbf{X}}(x)} - \sum_{x\in\mathbb{X},y\in\mathbb{Y}} P_{(\mathbf{X},\mathbf{Y})}(x,y)\log P_{\mathbf{Y}}(y) \\
&= \sum_{x\in\mathbb{X},y\in\mathbb{Y}} P_{\mathbf{X}}(x)P_{(\mathbf{Y}|\mathbf{X}=x)}(y)\log P_{(\mathbf{Y}|\mathbf{X}=x)}(y) - \sum_{x\in\mathbb{X},y\in\mathbb{Y}} P_{(\mathbf{X},\mathbf{Y})}(x,y)\log P_{\mathbf{Y}}(y) \\
&= \sum_{x\in\mathbb{X}} P_{\mathbf{X}}(x)\Big(\sum_{y\in\mathbb{Y}} P_{(\mathbf{Y}|\mathbf{X}=x)}(y)\log P_{(\mathbf{Y}|\mathbf{X}=x)}(y)\Big) - \sum_{y\in\mathbb{Y}}\Big(\sum_{x\in\mathbb{X}} P_{X,Y}(x,y)\Big)\log P_Y(y) \\
&= -\sum_{x\in\mathbb{X}} P_X(x)H(Y|X=x) - \sum_{y\in\mathbb{Y}} P_Y(y)\log P_Y(y) \\
&= -H(Y|X) + H(Y) \\
&= H(Y) - H(Y|X)
\end{aligned}
$$

3 Linear Algebra and Matrix Operations

3.1 Norms of a vector

Question: Find L_1, L_2 and L_∞ norms of the vector $\mathbf{v} = \begin{pmatrix} 0.5 \\ -3 \\ -1 \\ 2 \end{pmatrix}$

Solution:

L_p norm or p^{th} norm of a vector $V = \begin{pmatrix} v_1 \\ v_2 \\ . \\ . \\ . \\ v_n \end{pmatrix}$ is defined as

$$||V||_p := (\sum_{i=1}^{n} |v_i|^p)^{\frac{1}{p}}$$

Now, substituting $p = 1, 2, \infty$ gives L_1, L_2 and L_∞ norms respectively

$$||V||_1 = \sum_{i=1}^{n} |v_i|$$

$$||V||_2 = \sqrt{\sum_{i=1}^{n} v_i^2}$$

$$||V||_\infty = \max(|v_1|, |v_2|, ..., |v_n|) = \max_i |v_i|$$

For the given vector $V = \begin{pmatrix} 0.5 \\ -3 \\ -1 \\ 2 \end{pmatrix}$:

$$\begin{aligned} ||V||_1 &= \sum_{i=1}^{4} |v_i| \\ &= |0.5| + |-3| + |-1| + |2| \\ &= 0.5 + 3 + 1 + 2 \\ &= 6.5 \end{aligned}$$

$$\begin{aligned} ||V||_2 = \sqrt{\sum_{i=1}^{4} v_i^2} &= \sqrt{0.5^2 + (-3)^2 + (-1)^2 + 2^2} \\ &= \sqrt{0.25 + 9 + 1 + 4} \\ &= \sqrt{14.25} \\ &\approx 3.77 \end{aligned}$$

$$\begin{aligned} ||V||_\infty &= \max(|v_1|, |v_2|, ..., |v_4|) \\ &= \max(|0.5|, |-3|, |-1|, |2|) \\ &= \max(0.5, 3, 1, 2) \\ &= 3 \end{aligned}$$

3.2 Linear Independence, Rank

Question: Which of the following set of vectors are *linearly independent*?

$$\text{Set 1} = \left\{ \begin{pmatrix} 4 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 4 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 4 \end{pmatrix} \right\}$$

$$\text{Set 2} = \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right\}$$

$$\text{Set 3} = \left\{ \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 \\ 3 \\ 2 \end{pmatrix} \right\}$$

Solution:

A set of vectors $\{v_1, v_2, \ldots, v_n\}$ are said to be linearly independent if the only solution to the equation:

$$a_1 v_1 + a_2 v_2 + \cdots + a_n v_n = 0$$

is $a_1 = a_2 = \cdots = a_n = 0$. For set 1, consider:

$$a_1 \begin{pmatrix} 4 \\ 0 \\ 0 \end{pmatrix} + a_2 \begin{pmatrix} 0 \\ 4 \\ 0 \end{pmatrix} + a_3 \begin{pmatrix} 0 \\ 0 \\ 4 \end{pmatrix} = 0$$

$$4a_1 = 0$$

$$4a_2 = 0$$

$$4a_3 = 0$$

the only solution to the above system of equations is $a_1 = a_2 = a_3 = 0$. So, the vectors in set 1 are linearly independent.

For set 2, consider:

$$a_1 \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + a_2 \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} + a_3 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = 0$$

$$a_1 + a_2 + a_3 = 0$$

$$a_2 + a_3 = 0$$

$$a_3 = 0$$

solving the above system of equations give $a_1 = a_2 = a_3 = 0$. So, the vectors in set 2 are also linearly independent.

For set 2, consider:

$$a_1 \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix} + a_2 \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + a_3 \begin{pmatrix} 4 \\ 3 \\ 2 \end{pmatrix} = 0$$

$$2a_1 + 4a_3 = 0$$

$$a_1 + a2 + 3a_3 = 0$$

$$a_1 + 2a_3 = 0$$

Solving the above system of equations gives infinitely many solutions. One such solution is $a_1 = -2, a_2 = -1, a_3 = 1$. Since there is a non-zero solution, the vectors in set 3 are linearly dependent. Another way of looking at this problem is to see whether any vector in the set of vectors can be represented in terms of linear combination of other vectors. In Sets 1 and 2, all vectors are unique, and in Set 3, the third vector is simply the addition of first and second vectors making it redundant. That makes the entire set to be linearly dependent.

Question: What is the rank of the following matrices?

$$M_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, M_2 = \begin{pmatrix} 1 & 0 & 3 \\ 0 & 1 & 4 \end{pmatrix}, M_3 = \begin{pmatrix} 1 & 2 \\ 2 & 4 \\ 3 & 6 \end{pmatrix}$$

Solution:

One way of defining rank of a matrix is that the rank is equal to the minimum number of linearly independent rows or columns.

- For M_1, the number of linear independent rows equal to 2 and linearly independent columns equal to 2. So, rank$(M_1) = 2$.
- For M_2, the number of linear independent rows equal to 2 and linearly independent columns equal to 2. So, rank$(M_2) = 2$.
- For M_3, the number of linear independent rows equal to 1 and linearly independent columns equal to 1. So, rank$(M_3) = 1$.

3.3 Eigenvalues and Eigenvectors

Question: Find the eigenvalues and eigenvectors of the matrix: $M = \begin{pmatrix} -2 & 1 \\ 12 & -3 \end{pmatrix}$

Solution:

An eigenvector of a square matrix A is a non-zero vector x such that $Ax = \lambda x$ for some scalar λ. Here, the scalar λ is called eigenvalue of A corresponding to eigenvector x.

Using $Ax = \lambda x$, we have:

$$Ax - \lambda x = 0$$

$$Ax - \lambda I x = 0$$

$$(A - \lambda I)x = 0$$

Since eigenvector is non-zero, for the above equation to be true, determinant of $(A - \lambda I)$ has to be zero. We will get eigenvalues by equating the characteristic polynomial $det(A - \lambda I)$ to zero.

For the given matrix M, eigenvalues are the roots of the following equation:

$$\det(M - \lambda I) = 0$$

$$\det\left(\begin{pmatrix} -2 & 1 \\ 12 & -3 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) = 0$$

$$\det \begin{pmatrix} -2-\lambda & 1 \\ 12 & -3-\lambda \end{pmatrix} = 0$$
$$(-2-\lambda)(-3-\lambda) - 12 = 0$$
$$\lambda^2 + 5\lambda - 6 = 0$$
$$\lambda = 1, -6$$

where det is the determinant of a matrix. Now, substituting the eigenvalues in $(M - \lambda I)x = 0$ and solving for x gives eigenvectors.

If $\lambda = 1$, the equation becomes:

$$(M - I)x = 0$$
$$\begin{pmatrix} -3 & 1 \\ 12 & -4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0$$
$$-3x_1 + x_2 = 0$$
$$12x_1 - 4x_2 = 0$$

Solving the above two equation gives:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = t \begin{pmatrix} \frac{1}{3} \\ 1 \end{pmatrix}, t \in \mathbb{R}$$

If $\lambda = 6$, the equation becomes:

$$(M - I)x = 0$$
$$\begin{pmatrix} 4 & 1 \\ 12 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0$$
$$4x_1 + x_2 = 0$$
$$12x_1 + 3x_2 = 0$$

Solving the above two equations gives:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = t \begin{pmatrix} -\frac{1}{4} \\ 1 \end{pmatrix}, \quad t \in \mathbb{R}$$

Question: If x is a vector, then prove that $x^T x$ is the eigenvalue of the matrix xx^T corresponding to the eigenvector x.

Solution:

Consider the following:

$$\begin{aligned}(xx^T)x &= x(x^T x) \\ &= (x^T x)x\end{aligned}$$

This answers the question.

Question: Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of the matrix A corresponding to the eigenvectors $x_1, x_2, \ldots, x_n$. What are the eigenvalues of the matrix $A^2 + I$, where I is the identity matrix?

Solution:

It is given that:

$$Ax_i = \lambda_i x_i, \quad \forall i = 1, 2, \ldots, n$$

Consider $(A^2 + I)x_i$, we have:

$$\begin{aligned}(A^2 + I)x_i &= A^2 x_i + Ix_i \\ &= A(Ax_i) + x_i \\ &= A(\lambda_i x_i) + x_i \\ &= \lambda_i(Ax_i) + x_i \\ &= \lambda_i \lambda_i x_i + x_i \\ &= \lambda_i^2 x_i + x_i \\ &= (\lambda_i^2 + 1)x_i\end{aligned}$$

So, $\lambda_1^2 + 1, \lambda_2^2 + 1, \ldots, \lambda_n^2 + 1$ are the eigenvalues of the matrix $A^2 + I$ corresponding to the eigenvectors $x_1, x_2, \ldots, x_n$.

3.4 Singular Values and Singular Vectors

Question: Given a real matrix A of size $m \times n (m \geq n)$, what are its singular values and singular vectors, and how are they related to eigenvalues and eigenvectors of $A^T A$ and AA^T?

Solution:

Singular values σ_i of a matrix A are the square roots of non-zero eigen values λ_i of $A^T A$. Since $A^T A$ is positive semidefinite, all eigenvalues of $A^T A$ are non-negative. Non-zero eigenvalues of $A^T A$ and AA^T are the same. If v_i is an eigenvector of $A^T A$ corresponding to eigenvalue λ_i then Av_i is an eigenvector of AA^T corresponding to eigenvalue λ_i. This can be seen as follows.

$$(A^T A)v_i = \lambda_i v_i$$

Multiplying both sides by A:

$$A(A^T A)v_i = A\lambda_i v_i$$

$$(AA^T)(Av_i) = \lambda_i(Av_i)$$

$Av_i's$ are the fundamental set of eigenvectors of AA^T. n eigenvectors of $A^T A$ are called right singular vectors and m eigenvectors of AA^T are called left singular vectors of A. Eigenvectors v_i of $A^T A$ and eigenvectors u_i of AA^T are related as follows: since $A^T A$ is a symmetric matrix, its eigenvectors are orthogonal i.e., $v_i^T v_j = 0$ for $i \neq j$. Similarly, Av_j is orthogonal to Av_i, which implies:

$$(Av_j)^T(Av_i) = v_j^T(A^T Av_i) = v_j^T \sigma_i^2 v_i = \sigma_i^2(v_j^T v_i)$$

Assuming orthonormality:

$$Av_i = \sigma_i u_i$$

4 Numerical Methods

4.1 Taylor Series

Question: Write down the Taylor series expansion of a function $f(x)$ where $x \in \mathbb{R}$ and $f(x) \in \mathbb{R}$, and find the value of $e^{1.01}$ using a second-order approximation of its Taylor series.
Solution:

The Taylor series of a function $f(x)$ around a point x is defined as:

$$f(x+\Delta x) = f(x) + f'(x)\Delta x + \frac{f''(x)}{2!}(\Delta x)^2 + \cdots + \frac{f^n(x)}{n!}(\Delta x)^n + \dots$$

Using second-order approximation, we expand up to the second-order term to find the approximate value of a function. We need to find $f(1+0.01)$ where f is the exponential function.

$$\begin{aligned} f(1+0.01) &= f(1) + f'(1) \times 0.01 + \frac{f''(1)}{2!} \times (0.01)^2 \\ &= e + e \times 0.01 + \frac{e}{2} \times 0.0001 \\ &= e \times (1 + 0.01 + 0.00005) \\ &\approx 2.7456 \end{aligned}$$

Question: Write down the Taylor series of a function $f(\mathbf{x})$ where $\mathbf{x} \in \mathbb{R}^n$, $f(\mathbf{x}) \in \mathbb{R}$.
Solution:

The Taylor series of a function $f(\mathbf{x})$ around a point $\mathbf{x}$ is defined as,

$$f(\mathbf{x}+\mathbf{\Delta x}) = f(\mathbf{x}) + \nabla_{\mathbf{x}} f(\mathbf{x})\mathbf{\Delta x} + \frac{1}{2!}\mathbf{\Delta x}^T \nabla^{\mathbf{2}}_{\mathbf{x}} f(\mathbf{x})\mathbf{\Delta x} + \dots$$

where

$$\mathbf{\Delta x} = \begin{pmatrix} \Delta x_1 \\ . \\ . \\ . \\ \Delta x_n \end{pmatrix}$$

$$\nabla_{\mathbf{x}} \mathbf{f}(\mathbf{x}) = \begin{pmatrix} \frac{\partial \mathbf{x}}{\partial x_1} \\ . \\ . \\ . \\ \frac{\partial \mathbf{x}}{\partial x_n} \end{pmatrix}$$

$$\nabla^{\mathbf{2}}_{\mathbf{x}} \mathbf{f}(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 \mathbf{f}(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 \mathbf{f}(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 \mathbf{f}(\mathbf{x})}{\partial x_1 \partial x_n} \\ \frac{\partial^2 \mathbf{f}(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 \mathbf{f}(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 \mathbf{f}(\mathbf{x})}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \mathbf{f}(\mathbf{x})}{\partial x_n \partial x_1} & \frac{\partial^2 \mathbf{f}(\mathbf{x})}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 \mathbf{f}(\mathbf{x})}{\partial x_n^2} \end{pmatrix}$$

5 Basics of Machine Learning

5.1 L_1, L_2 regularization

Question: Why does the L_1 regularizer lead to sparse models and L_2 regularizer does not?
Solution:

Since the regularizer term can be thought of as a constraint which is added to the learning function of a machine learning model, the overall objective of the model is met at the point of contact of the graphs of learning objective function and the regularizer function. Due to the geometry of the L_1 norm (see figure below), it mostly meets the learning objective function at the corners of its graph where most indices are zero, leading to produce very low importance to respective features and thereby producing sparse models. For the L_2 regularizer, this is not the case, due to the geometry of the L_2 norm (see figure below). The regularizer's point of contact in this case with the learning objective is usually at a point where most indices are non-zero.

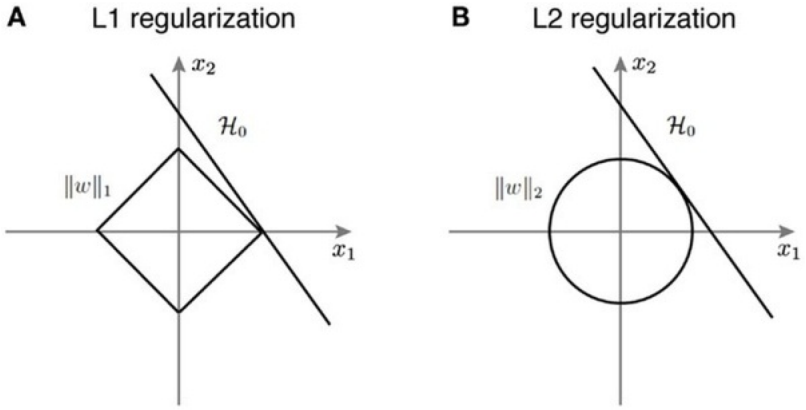


Figure 1: L1, L2 regularization

5.2 Mean Squared Error and Least Squares Regression

Question: State and find the derivative of Mean Squared Error for Least Squares Regression.
Solution:

The general objective of least squares regression in terms of mean square error is given as:

$$\min_w \frac{1}{2}\sum_{i=1}^{n}(w^T x_i - y_i)^2$$

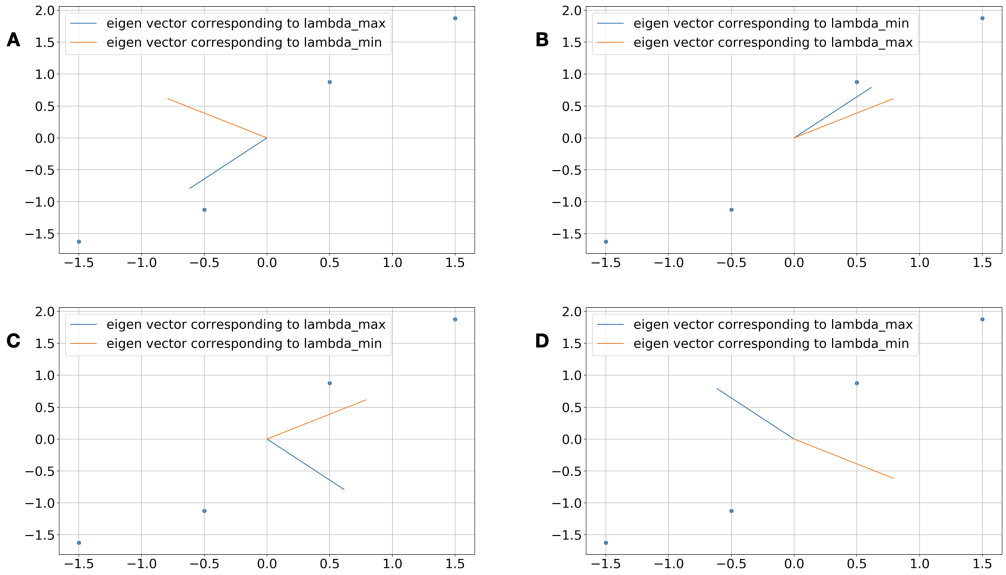
where x is an input data point and w are learnable parameters (which subsumes a bias parameter).

We take the derivative with respect to w_j as follows:

$$\begin{aligned}
\mathcal{L} &= \frac{1}{2}\sum_{i=1}^{n}(w^T x_i - y_i)^2 \\
\frac{\partial \mathcal{L}}{\partial w_j} &= \frac{1}{2} \times \frac{\partial \sum_{i=1}^{n}(w^T x_i - y_i)^2}{\partial w_j} \\
&= \frac{1}{2} \times \sum_{i=1}^{n} \frac{\partial (w^T x_i - y_i)^2}{\partial w_j} \\
&= \frac{1}{2} \times \sum_{i=1}^{n} 2 \times (w^T x_i - y_i) x_{ij} \\
&= \sum_{i=1}^{n} (w^T x_i - y_i) x_{ij}
\end{aligned}$$

5.3 Principal Component Analysis

Question: Suppose you are given a set of data points $P = \{(2,4), (1,3), (0,1), (-1, 0.5)\}$ and you need to write a program to find the direction of maximum variance in the dataset. Which of the following plots is likely to be your output? In the plots, lamba_max and lambda_min represents maximum and minimum eigenvalues of covariance matrix of dataset (Zoom into the plots if required).



Solution:

The direction of maximum variance in the dataset will be in the direction of eigenvector of covariance matrix of the dataset corresponding to the maximum eigenvalue. The given dataset clearly shows the maximum variance in one direction, and very low variance in another direction. In Plot A, the eigenvector corresponding to the maximum eigenvalue points in the direction of

maximum variance, and the other eigenvector corresponding to lowest eigenvalue points in the direction of lowest variance. So, Plot A is the correct output.

5.4 Logistic Regression

Question: Consider a binary classification problem. You have m data points in the form of (x_i, y_i) where x_i is i^{th} input point and $y_i \in \{0, 1\}$ is the class label and you are using the sigmoid function to predict the probability of the class based on weights w and input x. Write down the logistic loss based on the given setting.
Solution:

The logistic loss over the given data points is:

$$\mathcal{L} = \sum_{i=1}^{n} y_i \log P(y = 1) + (1 - y_i) \log(1 - P(y = 0)))$$

Since we are using sigmoid function as the probability function, we will get:

$$\mathcal{L} = \sum_{i=1}^{n} y_i \log(\frac{1}{1 + e^{-w^T x_i}}) + (1 - y_i) \log(1 - \frac{1}{1 + e^{-w^T x_i}})$$

5.5 Kernels & SVM

Question: Prove that $K(x, y) = x^T y$ is a valid kernel.
Solution:

A necessary and sufficient condition to check for validity of a function to be kernel is to check for positive semidefiniteness of the Gram matrix obtained from kernel function values as follows.

$$G_{i,j} = K(x_i, x_j) = x_i^T x_j$$

Now the entire Gram matrix G can be written as,

$$G = X^T X, \text{ where } X = (x_1, x_2, \ldots, x_n)^T$$

Now consider:

$$\begin{aligned} v^T G v &= v^T (X^T X) v \\ &= (v^T X^T)(Xv) \\ &= (Xv)^T (Xv) \\ &= ||Xv||_2^2 \geq 0 \end{aligned}$$

So, $K(x, y) = x^T y$ is a valid kernel.

5.6 Adaboost

Question: In the AdaBoost algorithm, if a decision stump (a decision tree of depth 1) misclassifies 1 out of 8 data points in one of the steps, what is the final importance of that stump in decision making?
Solution:

The decision stump misclassifies 1 out of 8 data points, i.e. its total error is $\frac{1}{8}$. In the AdaBoost algorithm, the importance value is computed as follows:

$$\begin{aligned} Importance &= \frac{1}{2}\log(\frac{1 - total\_error}{total\_error}) \\ &= \frac{1}{2}\log(\frac{1 - \frac{1}{8}}{\frac{1}{8}}) \\ &= \frac{1}{2}\log(7) \\ &\approx 0.97 \end{aligned}$$

5.7 Gaussian Mixture Models

Question: Which of the following statements about Gaussian Mixture Models is true?

1. Number of clusters $'k'$ in GMM is a learnable parameter.
2. GMM is a soft clustering technique.
3. Cluster assignment is done based on hard thresholding. That is, if the probability of a point belongs to a cluster is more than a specified threshold, that point is assigned to that cluster only.
4. GMM uses Expectation Maximization(EM) algorithm which makes GMM a Discriminative model.

Solution:

Learning a Gaussian Mixture Model is summarized as follows:

- Select predefined number of Gaussian clusters k and randomly assign mean and variance to those clusters.
- For each data point x_i, using Expectation Maximization (EM) algorithm, find the probability of that point belonging to a cluster $P(x_i|c_j)$ where c_j is cluster j. Repeat this for all clusters.
- Using Bayesian update, find the posterior probability of $P(c_j|x_i)$ and call it the weight of the cluster j given by point i.
- Using the above weights given by all data points for each cluster, perform weighted update to all cluster means and variances to get updated clusters.

From the above summary, we observe that k is predefined number of clusters that is not a learnable parameter. Since each point has some probability of belonging to a each cluster, it is a soft assignment rather than a hard assignment. Since we obtain the posterior probability using a Bayesian update to know the probability of cluster given a point, GMM can be considered a generative model. Correct answer is option 2.