2023-August-Mathematics of Network Algorithms

Problem Set
(Last Updated: November 6, 2023)

Linear Algebra

- Define the following terms (and every term used to define them) with an illustrative example:
  (a) Linear dependence and span of vectors (b) Norm of a vector
  (c) Eigenvalue, eigenvector and eigendecomposition
- Prove that ℓ_1, ℓ_2, and ℓ_∞ norm satisfies the properties mentioned while answering 1 (b).
- Write a 3×3 matrix $\mathbf{A}$ that is *not* identity, nor symmetric nor orthogonal. Also, write $\mathbf{A} \times \mathbf{A}$ and its transpose, inverse, determinant, eigenvalues, and eigenvectors.
- Write a (non-trivial) system of linear equations with at least 4 variables and 5 constraints both in equation form and matrix form.
- Consider a function $f : \mathbb{R}^2 \mapsto \mathbb{R}^2$ defined as $f((x_1, x_2)) = (-2x_2, -3x_1 + x_2)$. Describe geometrical interpretation of the above function in terms translation matrix, mention how a random point is shifted, and special points that may be only stretched.
- Consider a $n \times n$ matrix A. Prove that $\|A\|_F^2 = \mathrm{Tr}(A \cdot A^T)$.
- Prove that any symmetric matrix has real eigen values and the corresponding eigen vectors are orthogonal to each other.

Probability and Stastictics

- Describe the following terms with an illustrative examples:
  (a) Probability Space (b) Random variables
  (c) Frequentist probability and Bayesian probability
  (d) Mean, Variance, Covariance, & Correlation
- We flip a fair coin ten times. Find the probability of the following events: (i) Nr of heads and talks are equal. (ii) Nr of heads is more than nr of tails. (iii) The i^{th} flip and $(11-i)^{th}$ flip are same for every $i \in [5]$.
- We roll two fair dice. What is the probability space? What is the expectation of random variable representing the sum of two dice?
- Define the following distributions:
  (a) Bernoulli Distribution (b) Gaussian Distribution
  (c) Laplace Distributions (d) Multinoulli Distribution
  (e) Uniform Distribution
- Select your favourite distribution and derive expressions for its (i) expectation, (ii) variance, and (iii) standard deviation.

Numerical Optimization

- Consider the univariate function $f(x) = x^3 + 6x^2 - 3x - 5$. Find its stationary points and indicate whether they are maximum, minimum, or saddle points.
- Describe *overflow*, *underlow*, and *poor conditioning* with examples.
- Define *gradient* and *directional derivative*.
- Computer $\partial(f)/\partial\mathbf{x}$ when (i) $f = \sin(x_1)\cos(x_2)$, (ii) $f = 4x_1^2x_3 + 4x_1x^2x_3 + 5x_3^4$, and (iii) $f = x_1x_2x_4 + 2x_3^2x_4 + \sin(x_1x_2x_3)$.
- Prove the following identities:
  - $\partial(\mathbf{x}^\top\mathbf{x})/\partial\mathbf{x} = 2\mathbf{x}^\top$, $\partial(\mathbf{x}^\top\mathbf{a})/\partial\mathbf{x} = \mathbf{a}^\top$ and $\partial(\mathbf{a}^\top\mathbf{x})/\partial\mathbf{x} = \mathbf{a}^\top$
  - $\partial(\mathbf{a}^\top\mathbf{B}\mathbf{x})/\partial\mathbf{x} = \mathbf{B}^\top\mathbf{a}$, and $\partial(\mathbf{x}^\top\mathbf{B}\mathbf{x})/\partial\mathbf{x} = \mathbf{x}^T(\mathbf{B} + \mathbf{B}^\top)$
  - For symmetric matrix W, $\partial((\mathbf{x} - \mathbf{A}\mathbf{s})^\top\mathbf{W}(\mathbf{x} - \mathbf{A}\mathbf{s}))/\partial\mathbf{s} = -2(\mathbf{x} - \mathbf{A}\mathbf{s})^\top\mathbf{W}\mathbf{A}$.
- Prove that a function $f(\mathbf{x}) : \mathbb{R}^n \mapsto \mathbb{R}$ decreases fastest in the direction opposite to its gradient (assume that the gradient exists everywhere).
- Consider the optimization problem $\min\{\frac{1}{2}\mathbf{w}^\top\mathbf{w}\}$ over all $\mathbf{w} \in \mathbb{R}^n$ subjected to $\mathbf{w}^\top\mathbf{w} \geq 1$. Convert it into an unconstrained optimization problem by introducing Lagrange multiplier λ.
- Use gradient based optimisation to find $\mathbf{x}$ that minimizes $f(\mathbf{x}) = 1/2 \cdot ||\mathbf{A}\mathbf{x} \text{ - } \mathbf{b}||_2^2$.

Machine Learning Basics

- Describe the following terms with an illustrative examples:

  (a) Artificial Intelligence
  (b) Machine Learning
  (c) Deep Learning
  (d) Perceptron
  (e) Neural Network
  (f) Activation function
  (g) Loss Function
  (h) Optimisers
  (i) Parameters and Hyperparameters
  (j) Underfitting and Overfitting
  (k) Hypothesis space of a function
- Write steps in Principal Component Analysis to reduce 2-dimension data to 1-dimension data.
- Define *learning* in the context of Machine Learning.
- Write short description on five types of *tasks* (in the context of Machine Learning).
- What is *supervised learning* and *unsupervised learning*?
- Consider a learner regression problem where the objective is determine the value of $\mathbf{w} \in \mathbb{R}^n$ such that $\mathbf{w}^\top\mathbf{x}$ is as close to y as possible for vector $\mathbf{x}_i \in \mathbb{R}^n$ and scalar y_i for all $i \in [m]$. Derive an analytical expression to compute w if the difference between actual and computed values is determined using mean squared error.
- Describe *regularizer* with an example.

- Consider a set of samples $\{x^{(1)}, \ldots, x^{(m)}\}$ that are independently and indentically distributed according to a Bernoulli distribution with mean θ. Consider the following estimator $\hat{\theta}_m = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$. Compute `bias` and `variance` of the estimator.

- Consider a set of samples $\{x^{(1)}, \ldots, x^{(m)}\}$ that are independently and indentically distributed according to a Gaussian distribution with mean μ and variance σ^2. Compute `bias` and `variance` of the following estimators.
  - $\hat{\mu}_m = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$,
  - $\hat{\sigma}^2_m = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \hat{\mu}_m)^2$
  - $\hat{\sigma}^2_m = \frac{1}{m-1}\sum_{i=1}^{m}(x^{(i)} - \hat{\mu}_m)^2$

- Describe the following terms.
  - Cross-validation
  - Consistency of parameter estimation

- Compute maximum likelihood estimation of the relevant parameters for each of the following distribution.
  - Gaussian Distribution
  - Exponential Distribution
  - Geometric Distribution
  - Binomial Distribution
  - Poisson Distribution
  - Uniform Distribution

- Describe linear regression as maximum likelihood procedure.

- Describe (i) Artificial Intelligence, Machine Learning, and Deep Learning; (ii) Parameters and Hyperparameters; (iii) Underfitting and Overfitting; (iv) Hypothesis space of a function.

- Compute maximum likelihood estimation of the relevant parameters for Bernoulli Distribution.

- Design a Multilayer Perceptrons that determines if a list of length 4 is in sorted order, i.e, it receive four inputs x_1, x_2, x_3, x_4, where $x_i \in \mathbb{R}$, and outputs 1 if $x_1 < x_2 < x_3 < x_4$, and 0 otherwise. Only activation functions allowed are: Sigmoid, Step-Function, or ReLU.

- Let variable x can have values 1, 2 and 3 with probabilities $P(1) = 1/5$, $P(2) = 3/5$, and $P(3) = 1/5$. What is the expected value of x? Compare it with mean value of $(1, 2, 2, 2, 3)$?

- Explain *stochastic gradient descent* method and justify its use.

- We use wl^j_k to denote the weight for the connection from the k^{th} neuron in the $(\ell-1)^{\text{th}}$ layer to the j^{th} neuron in the ℓ^{th} layer, b^ℓ_j for the bias of the j^{th} neuron in the ℓ^{th} layer, and a^ℓ_j for the activation of the j^{th} neuron in the ℓ^{th} layer. Also, define

$$z^\ell_j := \sum_k w^\ell_{jk} a^{\ell-1}_k + b^\ell_k; \qquad a^\ell_j := \sigma(z^\ell_j); \qquad \delta^\ell_j := \frac{\partial C}{\partial z^\ell_j},$$

where σ is a sigmoid function. Prove the following equations:

$$(1)\ \delta^L_j = \frac{\partial C}{\partial a^L_j}\sigma'(z^L_j), \quad (2)\ \delta^\ell = [[w^{\ell+1}]^T\delta^{\ell+1}] \odot \sigma'(z^\ell), \quad (3)\ \frac{\partial C}{\partial b^\ell_j} = \delta^\ell_j, \quad (4)\ \frac{\partial C}{\partial w^\ell_{jk}} = a^{\ell-1}_k \delta^\ell_j.$$

For (1), L is the index of the output layer.

- Write a back-propagation algorithm with stochastic gradient descent when the activation function is sigmoid and the cost function is mean squared error.
- Write back-propagation algorithm when the activation function is a linear function $\sigma(z) = 2z$, and the cost function is mean squared error.
- Write the two assumptions we make about the cost function and justify them.
- Define *cross-entropy cost function* and demonstrate that it is useful in the case of binary classification.
- Justify why *cross-entroy cost function* might be more useful than *quadratic cost function*.
- Consider the sigmoid function $\sigma(z)$. Prove that $\sigma'(z) = \sigma(z)(1 - \sigma(z))$.
- Define *soft-max activation function*. Mention the cases in which this activation function is useful. Justify the usefulness over sigmoid activation function.
- Define *regularisation*. Briefly explained at least three techniques for it.
- Consider a cost function C_0. Define C as a cost function obtained by adding a weighted decay term for the regularisation. Derive terms to modify weights and biases with this modification.
- Derive expressions to update weights and biases while using stochastic gradient descent when cost function is modified with L_2-regularization.
- Explain, with an illustrative example, use of regularization to reduce overfitting.
- Define *dropouts* and justify its use.
- Define *weighted initialization* and justify its use when the activation function is a sigmoid.
- Consider a continious function $f : \mathbb{R}^d \mapsto \mathbb{R}$. Justify that there is artificial neural network that approximates the function f.
- What is the *vanishing gradient problem* and why it occur? Explain it using simple artificial neural network.
- Define *convolution neural networks*. Specify a toy example that highlights shared weights and biases, and pooling layer.

- Define following terms: ∘ Vectorization, ∘ k-fold Validation, ∘ Iterated k-fold validation, ∘ Hold-out validation, and ∘ Feature engineering
- Describe a flow of designing ML/DL algorithm to the best of your abilities.