

Exercises

Machine Learning: Foundations and Applications
MATH CS 120AI

Paul J. Atzberger
<https://web.atzberger.org/>

Can choose to complete 1 out the following 3 problems.

- (Neural Network Universal Approximation and Activations) The Cybenko Theorem states that if a continuous activation function $g(z)$ is discriminatory on the unit cube $I_n \subset \mathbb{R}^n$ then the linear space $\mathcal{V} = \{q \mid q(\mathbf{x}) = \sum_{j=1}^n \alpha_j g(\mathbf{w}_j^T \mathbf{x} + b_j), n \in \mathbb{N}\}$ is dense in the space of continuous functions $\mathcal{C}(I_n)$. In other words, for any continuous function $f \in \mathcal{C}(I_n)$ and $\epsilon > 0$, there exists a $q \in \mathcal{V}$ such that $|f(\mathbf{x}) - q(\mathbf{x})| < \epsilon$ for all $\mathbf{x} \in I_n$. An activation function $g(z)$ is said to be discriminatory if for a Borel measure $\mu \in \mathcal{M}$ we have for all weights $\mathbf{w}, b$ that $\int g(\mathbf{w}^T \mathbf{x} + b) d\mu(\mathbf{x}) = 0$ then the measure must be zero $\mu \equiv 0$.
 - Show that the sigmoid activation function $g(z) = 1/(1 + e^{-z})$ is discriminatory on $I_1 = [0, 1]$. Hint: Use that $\int g(\mathbf{w}^T \mathbf{x} + b) d\mu(\mathbf{x}) = 0$ for all $\mathbf{w}, b$ iff $\int q(\mathbf{x}) d\mu(\mathbf{x}) = 0$ for all $q \in \mathcal{V}$.
 - Show that the ReLU activation function $g(z) = \max(z, 0)$ is discriminatory on I_1 . Hint: Use that $\int g(\mathbf{w}^T \mathbf{x} + b) d\mu(\mathbf{x}) = 0$ for all $\mathbf{w}, b$ iff $\int q(\mathbf{x}) d\mu(\mathbf{x}) = 0$ for all $q \in \mathcal{V}$.
 - Show that the linear activation function $g(z) = z$ is not discriminatory on I_1 . Hint: Construct a counter-example using a measure of the form $\mu(x) = a_1 \delta(x - x_1) + a_2 \delta(x - x_2) + a_3 \delta(x - x_3)$, where $\delta(\cdot)$ denotes here the Dirac δ -function (measure).
- (Approximating Distributions / Variational Inference) Consider a data distribution of the form $z_i \sim \bar{\rho}$ where $\bar{\rho}(z) = (1/Z) \exp(-\beta U(z))$, where $U(z) = (1 - z^2)^2 - 4\alpha z$, $Z = \int \exp(-\beta U(z))$, $\alpha = 0.171 = 9/10 - (9/10)^3$.
 - Plot the distribution for the range $z \in [-3, 3]$ when $\beta = 1.5$.
 - Show the critical points $U'(z) = 0$ are at $z_- = -9/10$, $z_+ = \frac{1}{20}(9 + \sqrt{157})$ and $z_0 = \frac{1}{20}(9 - \sqrt{157})$. The $z_{\pm}$ give local minima of U , and consequently local maxima of $\bar{\rho}$.
 - Suppose as a generative model a Gaussian is used with parameter $\mu = z_+$ and $\sigma^2 = (\beta u''(z_+))^{-1}$, denote this distribution by $\tilde{\rho}$. Derive this Gaussian estimate for $\beta \gg 1$ by computing Taylor expansion of $U(z) = U(z_+) + U'(z_+)(z - z_+) + \frac{1}{2}U''(z_+)(z - z_+)^2 + \dots + R(z, z_+)$ and substituting for U . The approximation $\tilde{\rho}$ of $\bar{\rho}$ is obtained by retaining the leading order terms in $\beta(z - z_+)^{\ell}$ with $\ell \leq 2$.
 - How well does $\tilde{\rho}$ approximate the distribution $\bar{\rho}$? For $\beta = 0.5, 1.0, 1.5, 2.0, 3.0$ compute numerical estimates of the L^1 -norm error $\|\tilde{\rho} - \bar{\rho}\|_1$.
- (Generative Models and MLE) The method of Maximum Likelihood Estimation (MLE) selects a model θ^* from data $\{\mathbf{z}_i\}_{i=1}^m$ by optimizing

$$\theta^* = \arg \min_{\theta \in \Theta} \sum_{i=1}^m -\log(\rho(\mathbf{z}_i; \theta)).$$

- Consider estimating the mean and variance of a Gaussian using MLE. Show that the MLE estimate $\hat{\theta}$ for the variance σ^2 of a Gaussian $z_i \sim \eta(\mu, \sigma^2)$ is biased. An estimator

$\hat{\theta}$ is called biased if $\mathbb{E}[\hat{\theta}] - \sigma^2 \neq 0$, where σ^2 is the true parameter value of the data distribution.

- (b) Consider using linear models for regression $\mathcal{H} = \{\mathbf{w}^T \mathbf{x} + b \mid \mathbf{w} \in \mathbb{R}^n, b \in \mathbb{R}\}$, where $\theta = (\mathbf{w}, b)$ and $\mathbf{z}_i = (\mathbf{x}_i, \mathbf{y}_i)$. Assume a Gaussian noise model for the data generation $y_i = f(\mathbf{x}_i) + \xi_i$ with i.i.d. $\xi_i \sim \eta(0, \sigma)$, $f \in \mathcal{H}$, and $n = 1$. Show in this case that MLE is equivalent to performing least-squares regression.
- (c) Use MLE to find the estimate for $\theta = (\mu, \sigma^2)$ in the limit of an infinite number of samples (i.e. replace the MLE objective sum with $\int -\log(\rho(z; \theta)) \bar{\rho}(z) dz$). Show that for Gaussian generative models $\rho(z; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} (z - \mu)^2\right)$ the infinite sample MLE approximates $\bar{\rho}$ by selecting for $\rho(z; \theta)$ the parameters $\mu = \int z \bar{\rho}(z) dz$ and $\sigma^2 = \int (z - \mu)^2 \bar{\rho}(z) dz$.