MATH 173B, HW2

Victor Pekkari — epekkari@ucsd.edu

February 1, 2025

Problem 1

(a)

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx.$$

Substitute: $x = (x - \mu) + \mu$:

$$E(X) = \int_{-\infty}^{\infty} [(x - \mu) + \mu] f_X(x)\, dx.$$

$$E(X) = \int_{-\infty}^{\infty} (x - \mu) f_X(x)\, dx + \int_{-\infty}^{\infty} \mu f_X(x)\, dx.$$

The first integral evaluates to zero since $(x - \mu)$ is odd and $f_X(x)$ is even:

$$\int_{-\infty}^{\infty} (x - \mu) f_X(x)\, dx = 0.$$

$$\mu \int_{-\infty}^{\infty} f_X(x)\, dx = \mu \cdot 1 = \mu.$$

Thus, we have shown that:

$$E(X) = \mu.$$

(b)

$$\mathrm{Var}(X) = E(X^2) - (E(X))^2.$$

From the expectation proof, we know $E(X) = \mu$

$$\mathrm{Var}(X) = E(X^2) - \mu^2.$$

The expectation $E(X^2)$ is:

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f_X(x)\, dx.$$

Rewriting x^2 as $(x-\mu)^2 + 2\mu(x-\mu) + \mu^2$, we get:

$$E(X^2) = \int_{-\infty}^{\infty} [(x-\mu)^2 + 2\mu(x-\mu) + \mu^2] f_X(x)\, dx.$$

$$E(X^2) = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)\, dx + 2\mu \int_{-\infty}^{\infty} (x-\mu) f_X(x)\, dx + \mu^2 \int_{-\infty}^{\infty} f_X(x)\, dx.$$

The second integral is zero (as shown in the expectation proof), and the last integral evaluates to 1, so:

$$E(X^2) = \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x)\, dx + \mu^2 = \sigma^2 \cdot \int_{-\infty}^{\infty} f_X(x)\, dx + \mu^2$$

(By definition, $\sigma = x - \mu$)

$$E(X^2) = \sigma^2 + \mu^2.$$

Substituting into the variance formula:

$$\mathrm{Var}(X) = (\sigma^2 + \mu^2) - \mu^2 = \sigma^2.$$

Problem 2

(a)

A valid CDF must satisfy:

(1) Monotonicity: $G_Z(z)$ must be non-decreasing.

$$G'_Z(z) = \frac{e^{-z}}{(1+e^{-z})^2} > 0, \quad \forall z \in \mathbb{R}.$$

(2) Limits: the $G(Z) \xrightarrow[z\to\infty]{} 1$ and $G(Z) \xrightarrow[z\to-\infty]{} 0$

$$\lim_{z\to-\infty} G_Z(z) = \frac{1}{1+e^x} = 0, \quad \lim_{z\to\infty} G_Z(z) = \frac{1}{1+e^{-x}} = 1.$$

(3) Right-continuity: A function that is continous is also right-continous:

A function $f(x)$ is continuous on the set Ω at a point $x = a$ if:

$$\lim_{x\to a} f(x) = f(a) \quad \forall a \in \Omega \tag{1}$$

To check the continuity of $\sigma(x)$, we evaluate:

$$\lim_{x\to a} \frac{1}{1+e^{-x}} \tag{2}$$

The exponential function e^{-x} is well defined for all real numbers, and the denominator is never going to be zero. We can therefore conclude that the sigmoid-function is contious for all real numbers.

it follows that:

$$\lim_{x \to a} \sigma(x) = \sigma(a), \quad \forall a \in \mathbb{R}.$$

Conclusion: Since everything holds for $G_Z(z)$, it is a valid CDF

(b)

$y = 1$ when $z \geq -a$

$y = -1$ when $z < -a$

Using the definition of a CDF:

$$\mathbb{P}(y = 1) = \mathbb{P}(z \geq -a) = 1 - G_Z(-a)$$

$$\mathbb{P}(y = -1) = G_Z(-a).$$

Since $G_Z(-a) = \frac{1}{1+e^a}$, we get:
Answer:

$$\mathbb{P}(y = 1) = \frac{e^a}{1 + e^a}$$

$$\mathbb{P}(y = -1) = \frac{1}{1 + e^a}$$

(c)

From part (b), we can express the probabilities as:

$$\mathbb{P}(y = 1) = \frac{1}{1 + e^{-a}}, \quad \mathbb{P}(y = -1) = \frac{1}{1 + e^a}.$$

For general $y \in \{-1, 1\}$ we can combine the two functions above as follows:

$$p(y) = \frac{1}{1 + e^{-ay}}$$

(d)

We can rewrite the joint distribution as the following since $y_1, ..., y_n$ are independent:

$$p(y_1, y_2, \ldots, y_N) = \prod_{i=1}^{N} p(y_i) = \prod_{i=1}^{N} \frac{1}{1 + e^{-a_i y_i}}.$$

Problem 3

(a)

Answer: It makes sense to maximize $H(w)$ since by maximizing $H(w)$ we want to maximize every term $\frac{1}{1+e^{-w^T x_i y_i}}$ in the product notation $\prod_{i=1}^{N}$.

To maximize the terms $\frac{1}{1+e^{-w^T x_i y_i}}$ we have to minimize their denominator which mean we want $e^{-w^T x_i y_i}$ to be as small as possible. To keep it small we want to keep the exponent of e negative, $-w^T x_i y_i < 0$ or:

$$w^T x_i y_i > 0 \tag{*}$$

By viewing x_i as different inputs and y_i as our different targets. It is clear that to keep (*) greater than 0 we have to make :

$$sign(w^T x_i) = sign(y_i)$$

and we also want the magnitude of: w^T to be as big as possible. This means that for every training example x_i, the decision boundary defined by w^T should correctly classify y_i by ensuring that $w^T x_i$ has the same sign as y_i. When this condition holds, the probability assigned to the correct class label is high.
Additionally, maximizing $w^T x_i y_i$ not only ensures correct classification but also increases confidence in the prediction. A larger magnitude of $w^T x_i$ results in values closer to 1 in the sigmoid function:

$$\frac{1}{1+e^{-w^T x_i y_i}}$$

pushing the probability of correct classification towards 1.

Since $H(w)$ is the product of these probabilities across all training samples, maximizing $H(w)$ is equivalent to maximizing the likelihood of the observed labels under the logistic regression model. This aligns with the principle of maximum likelihood estimation (MLE), which seeks to find the parameter w that best explains the observed data.

(b)

$$log\left[H(w)\right] = \log\left[\prod_{i=1}^{N} \frac{1}{1+e^{-w^T x_i y_i}}\right]$$

$$= \sum_{i=1}^{N} \log\left[\frac{1}{1+e^{-w^T x_i y_i}}\right]$$

(c)

Maximizing the LHS side is the same as minimizing the RHS, and minimizing the RHS is the same as minimizing $F(w)$ since the only difference is a factor of $\frac{1}{N}$

$$\sum_{i=1}^{N} \log\left[\frac{1}{1+e^{-w^T x_i y_i}}\right] = -\sum_{i=1}^{N} \log\left[1+e^{-w^T x_i y_i}\right]$$

(d)

Answer: To derive an SGD algorithm for minimizing $F(w)$, we first derive the gradient of the objective function:

$$\nabla F(w) = \frac{1}{N}\sum_{i=1}^{N} \nabla \log\left(1+e^{-w^T x_i y_i}\right).$$

The gradient of the individual term is:

$$\nabla \log\left(1+e^{-w^T x_i y_i}\right) = -\frac{y_i x_i e^{-w^T x_i y_i}}{1+e^{-w^T x_i y_i}} = -y_i x_i \left(1-\frac{1}{1+e^{-w^T x_i y_i}}\right).$$

Below is the stochastic gradient descent update step, using index i_t uniformly independently chosen from $\{1, 2, ...N\}$ where N is the number of points. η is the constant step size.
(A constant step size in SGD won't make us converge fully to an optimizer)

$$w_{t+1} = w_t - \eta \nabla f_{i_t}(w_t),$$

where

$$f_{i_t}(w_t) = \log\left(1+e^{-w^T x_i y_i}\right)$$

$$\implies \nabla f_{i_t}(w_t) = -y_{i_t} x_{i_t} \left(1-\frac{1}{1+e^{-w_t^T x_{i_t} y_{i_t}}}\right).$$

(e)

$$F(w) = \sum_{i=1}^{N} f_{i_t}(w) \qquad (*)$$

$$(*) \implies \mathbb{E}\nabla f_{i_t}(w) = \frac{1}{N}\cdot\sum_{i=1}^{N} \nabla f_{i_t}(w) = \nabla F(w)$$

$F(w)$ and f_{i_t} as computed in (d):

$$\nabla F(w) = \sum_{i=1}^{N} -y_{i_t} x_{i_t} \left(1 - \frac{1}{1 + e^{-w_t^T x_{i_t} y_{i_t}}}\right)$$

$$\nabla f_{i_t}(w_t) = -y_{i_t} x_{i_t} \left(1 - \frac{1}{1 + e^{-w_t^T x_{i_t} y_{i_t}}}\right)$$