Discussion 3

Note: *Your TA will probably not cover all the problems on this worksheet. The discussion worksheets are not designed to be finished within an hour. They are deliberately made slightly longer so they can serve as resources you can use to practice, reinforce, and build upon concepts discussed in lectures, discussions, and homework.*

This Week's Cool AI Demo/Video:
`https://youaretheassistantnow.com/`

`https://gemini.google.com/`

1 MLE with a Linear Constraint on Independent Gaussians

(a) Let $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ be independent. You only observe their sum

$$S := X + Y.$$

Find the Maximum Likelihood Estimate (MLE) of X given $S = 2$.

(i) Show that, up to a proportionality constant,

$$p_{X|S}(x \mid 2) \propto p_X(x)\, p_Y(2 - x).$$

(ii) Use the Gaussian forms to show that maximizing $p_{X|S}(x \mid 2)$ with respect to x is equivalent to minimizing

$$\frac{(x - \mu_X)^2}{2\sigma_X^2} + \frac{(2 - x - \mu_Y)^2}{2\sigma_Y^2}.$$

(iii) Find the estimate, $\hat{x}_{\text{MLE}}$, that maximizes the likelihood $p_{X|S}(x \mid 2)$ using the expression you found in (ii). Simplify to a weighted-average form.

(iv) Consider the special case where $\mu_X = \mu_Y = 0$ and $\sigma_X^2 = \sigma_Y^2$. How can we simplify the maximum likelihood estimate?

Solution:

(i) Using the definition of conditional probability for continuous random variables,

$$p_{X|S}(x \mid s) = \frac{p_{X,S}(x, s)}{p_S(s)}.$$

Because S is the sum of two independent random variables,

$$p_{X,S}(x, s) = p_{X,Y}(x, s - x) = p_X(x) p_Y(s - x).$$

Combining this with the previous statement,

$$p_{X|S}(x \mid s) = \frac{p_X(x)p_Y(s-x)}{p_S(s)}.$$

The density $p_S(s)$ does not depend on x, so

$$p_{X|S}(x \mid 2) \propto p_X(x)\, p_Y(2-x).$$

(ii) First, recall the density function for a Gaussian distribution:

$$p_X(x) = \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left(-\frac{(x-\mu_X)^2}{2\sigma_X^2}\right) \propto \exp\left(-\frac{(x-\mu_X)^2}{2\sigma_X^2}\right).$$

Using this definition and properties of the exponential function,

$$p_{X|S}(x \mid 2) \propto p_X(x)\, p_Y(2-x) \propto \exp\left(-\frac{(x-\mu_X)^2}{2\sigma_X^2} - \frac{(2-x-\mu_Y)^2}{2\sigma_Y^2}\right).$$

Because the exponential function is monotonically increasing, maximizing $p_{X|S}(x \mid 2)$ with respect to x is the same as maximizing the term inside the exponential. This is then equivalent to minimizing the negation of this term.

(iii) As we showed in the previous parts, to find the value of x that maximizes $p_{X|S}(x \mid 2)$, we can find the value that minimizes:

$$\frac{(x-\mu_X)^2}{2\sigma_X^2} + \frac{(2-x-\mu_Y)^2}{2\sigma_Y^2}.$$

Because this is a quadratic function of x, we can find the minimizer by taking the derivative and setting it equal to zero:

$$\frac{d}{dx}\left[\frac{(x-\mu_X)^2}{2\sigma_X^2} + \frac{(2-x-\mu_Y)^2}{2\sigma_Y^2}\right] = \frac{x-\mu_X}{\sigma_X^2} - \frac{2-x-\mu_Y}{\sigma_Y^2} = 0.$$

Solving for x,

$$x\left(\frac{1}{\sigma_X^2} + \frac{1}{\sigma_Y^2}\right) = \frac{\mu_X}{\sigma_X^2} + \frac{2-\mu_Y}{\sigma_Y^2}.$$

Hence, the maximum likelihood estimate for x is

$$\hat{x}_{\text{MLE}} = \frac{\frac{\mu_X}{\sigma_X^2} + \frac{2-\mu_Y}{\sigma_Y^2}}{\frac{1}{\sigma_X^2} + \frac{1}{\sigma_Y^2}} = \boxed{\frac{\sigma_Y^2\mu_X + \sigma_X^2(2-\mu_Y)}{\sigma_X^2 + \sigma_Y^2}.}$$

(iv) If $\mu_X = \mu_Y = 0$ and $\sigma_X^2 = \sigma_Y^2$, then

$$\hat{x}_{\text{MLE}} = \frac{2}{2} = \boxed{1}.$$

2 Binomial MLE with Misclassification

(a) You run m independent experiments. In each experiment, you perform n independent Bernoulli trials with unknown true success probability p. However, your measurement device is imperfect:

- A true success is recorded as a success with probability $1-q$ (and as a failure with probability q).
- A true failure is recorded as a success with probability r (and as a failure with probability $1-r$).

Let $X_i \sim \mathrm{Bin}(n,p)$ denote the unobserved number of true successes in experiment i, and let Y_i be the observed number of recorded successes. Recall that the probability mass function for the binomial random variable $X_i \sim \mathrm{Bin}(n,p)$ is

$$p(x_i) = \binom{n}{x_i} p^{x_i} \left(1-p\right)^{n-x_i}.$$

Derive the likelihood for the observations $\mathcal{D} = \{y_i\}_{i=1}^m$. *Hint: Your answer should be a function of the true success probability p.*

Solution:

The variable Y_i is also a binomial random variable, where n is the number of trials, but the probability of an observed success is now a function of p:

$$\pi(p) = (1-q)p + r(1-p) = r + (1-q-r)p.$$

Therefore, we can express the probability mass function parameterized by p:

$$p(y_i|p) = \binom{n}{y_i} \pi(p)^{y_i} \left(1-\pi(p)\right)^{n-y_i}.$$

Since experiments are independent, the likelihood of our full dataset

$$p(\mathcal{D}|p) = \prod_{i=1}^m \binom{n}{y_i} \pi(p)^{y_i} \left(1-\pi(p)\right)^{n-y_i}.$$

(b) Find the MLE $\widehat{p}$ in closed form. Show all steps.

Solution:

We want to find an expression for p that maximizes $p(\mathcal{D}|p)$. Let's begin with simplifying this likelihood by defining $s := \sum_{i=1}^m y_i$ and writing

$$p(\mathcal{D}|p) = \left(\prod_{i=1}^m \binom{n}{y_i}\right) \pi(p)^s \left(1-\pi(p)\right)^{mn-s}.$$

Notice that the first product does not depend on p, so we have

$$p(\mathcal{D}|p) \propto \pi(p)^s \left(1-\pi(p)\right)^{mn-s}.$$

Now, as we often do in MLE, we will look at the log likelihood:

$$\ln p(\mathcal{D}|p) \propto s \ln \pi(p) + (mn - s) \ln \big(1 - \pi(p)\big).$$

To find that expression for p that maximizes the log likelihood, notice that $\pi(p)$ is a linear function of p and the likelihood is a concave function of $\pi(p)$. Therefore, we can find the maximizer by taking the derivative with respect to p using the chain rule:

$$\frac{\partial}{\partial p} \ln p(\mathcal{D}|p) = \frac{\partial}{\partial \pi(p)} \ln p(\mathcal{D}|p) \frac{\partial}{\partial p} \pi(p).$$

Computing each factor in this product,

$$\frac{\partial}{\partial \pi(p)} \ln p(\mathcal{D}|p) = \frac{s}{\pi(p)} - \frac{mn - s}{1 - \pi(p)}, \qquad \frac{\partial}{\partial p} \pi(p) = 1 - q - r.$$

Therefore,

$$\frac{\partial}{\partial p} \ln p(\mathcal{D}|p) = \left(\frac{s}{\pi(p)} - \frac{mn - s}{1 - \pi(p)} \right) (1 - q - r).$$

Now we need to set this expression equal to zero and solve for the critical points:

$$\frac{\partial}{\partial p} \ln p(\mathcal{D}|p) = 0 \quad \Longrightarrow \quad \frac{s}{\pi(p)} - \frac{mn - s}{1 - \pi(p)} = 0 \quad \Longrightarrow$$

$$s(1 - \pi(p)) = (mn - s)\pi(p) \Longrightarrow \quad s = mn\pi(p) \quad \Longrightarrow \quad \pi(p) = \frac{s}{mn}$$

Now let's plug in our expression for $\pi(p)$ so we can we can solve for p:

$$r + (1 - q - r)p = \frac{s}{mn} \quad \Longrightarrow \quad \boxed{= \frac{\frac{S}{mn} - r}{1 - q - r}.}$$

3 Proof of K-means Convergence

Consider the K-means algorithm applied to data points $\{x_n\}_{n=1}^N \subset \mathbb{R}^D$ with K clusters. Define binary indicator variables $r_{nk} \in \{0, 1\}$ with $\sum_{k=1}^K r_{nk} = 1$ for each n, and cluster centers $\{\mu_k\}_{k=1}^K$. The K-means objective is

$$J(\{r_{nk}\}, \{\mu_k\}) \;=\; \sum_{n=1}^N \sum_{k=1}^K r_{nk} \, \|x_n - \mu_k\|^2.$$

The K-means algorithm alternates between

1. **Update assignments**: Set $r_{nk} = 1$ for the k that minimizes $\|x_n - \mu_k\|^2$ and $r_{nj} = 0$ for all $j \neq k$.
2. **Update means**: Set the cluster centers to

$$\mu_k \;=\; \frac{\sum_{n=1}^N r_{nk} x_n}{\sum_{n=1}^N r_{nk}} \qquad \text{(for clusters with } \sum_n r_{nk} > 0\text{)}.$$

Prove that the K-means algorithm converges after a finite number of iterations. (Hint: It is sufficient to show that the objective is monotonically decreasing, and that there are a finite number of iterations).

Solution:

We show two things: (A) each assignment or update step does not increase the objective J and decreases it unless a fixed point is reached; (B) there are only finitely many possible assignments $\{r_{nk}\}$, so decrease steps cannot occur forever — hence the algorithm must terminate after finitely many iterations.

(A) The objective never increases.

- *Assignment step decreases (or leaves unchanged) J:* Fix the current centers $\{\mu_k\}$. For each data point x_n, choosing $r_{nk} = 1$ for the nearest center minimizes its contribution to J. Hence J does not increase.
- *Update step decreases (or leaves unchanged) J:* Fix $\{r_{nk}\}$. The objective for each cluster k is

  $$J_k(\mu) = \sum_{n: r_{nk}=1} \|x_n - \mu\|^2,$$

  minimized uniquely by the mean of the assigned points as shown in lecture:

  $$\mu_k^\star = \frac{\sum_{n=1}^N r_{nk} x_n}{\sum_{n=1}^N r_{nk}}.$$

  Hence updating μ_k to the mean does not increase J.

(B) Only finitely many assignments exist. Each point has K choices independently, so there are at most K^N assignment matrices.

Combine (A) and (B). Since J decreases strictly whenever assignments or centers change, and there are only finitely many assignments, the algorithm must eventually reach a state where neither step changes anything. At that point, K-means has converged after a finite number of iterations.