CS 189 Introduction to Machine Learning
Spring 2022 Marvin Zhang

DIS1

1 Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) is a method of estimating the parameters of a statistical model given observations, by finding the parameters that maximize the likelihood of the observations. Concretely, given observations $y_1, y_2, \ldots, y_n$ distributed according to $p_\theta(y_1, y_2, \ldots, y_n)$ (here p_θ can be a probability mass function for discrete observations or a density for continuous observations), the likelihood function is defined as $L(\theta) = p_\theta(y_1, y_2, \ldots, y_n)$ and the MLE is

$$\hat{\theta}_{\text{MLE}} = \arg\max_\theta L(\theta).$$

We often make the assumption that the observations are *independent and identically distributed* or iid, in which case $p_\theta(y_1, y_2, \ldots, y_n) = p_\theta(y_1) \cdot p_\theta(y_2) \cdot \cdots \cdot p_\theta(y_n)$.

(a) Your friendly TA recommends maximizing the log-likelihood $\ell(\theta) = \log L(\theta)$ instead of $L(\theta)$. **Why does this yield the same solution $\hat{\theta}_{\text{MLE}}$? Why is it easier to solve the optimization problem for $\ell(\theta)$ in the iid case? Write down both $L(\theta)$ and $\ell(\theta)$ for the Gaussian $f_\theta(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y-\mu)^2}{2\sigma^2}}$ with $\theta = (\mu, \sigma)$.**

Solution: As the log is strictly monotonically increasing, maximizing $\ell(\theta) = L(\theta)$ and $L(\theta)$ will yield the same solution. Concretely, if θ^* is a unique maximum of $L(\theta)$, we have $L(\theta) < L(\theta^*)$ for all $\theta \neq \theta^*$ in the parameter space and therefore due to strict monotonicity of the log, $\ell(\theta) = \log L(\theta) < \log L(\theta^*) = \ell(\theta^*)$, which means θ^* is also a unique maximum of $\ell(\theta)$.

In the iid case, the log-likelihood decomposes into a sum

$$\ell(\theta) = \sum_{i=1}^{n} \log f_\theta(y_i)$$

and it is often easier to optimize over these sums rather than products:

Numerically: There are special algorithms like stochastic gradient descent available for sums that you will learn about later in lecture. Another reason is that forming the product of many probabilities will yield a very small number and it is easy to generate a floating point underflow this way. On the other hand, adding the logs of probabilities is a more stable operation because the partial sums stay in a reasonable range.

Analytically: Usually it is easier to compute the gradient of $\ell(\theta)$ than for $L(\theta)$. As an example, consider the case of a Gaussian distribution:

The likelihood function is

$$L(\theta) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \cdot e^{-\frac{\sum_{i=1}^{n}(y_i-\mu)^2}{2\sigma^2}}.$$

Taking logs yields

$$\ell(\theta) = \sum_{i=1}^{n} \log f_\theta(y_i) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2$$

which is much easier to minimize than $L(\theta)$.

(b) **What is $\int p_\theta(y_1, y_2, \ldots, y_n)\, dy_1 \cdots dy_n$? Can we say anything about $\int p_\theta(y_1, y_2, \ldots, y_n) d\theta$?**

Solution: The main point of this problem is to highlight that even though the likelihood function may be comprised of probabilities or probability densities, it does not generally form a valid probability distribution over the parameter.

The probability distribution is normalized, therefore by integrating over the observations we have

$$\int p_\theta(y_1, y_2, \ldots, y_n)\, dy_1 \cdots dy_n = 1.$$

In terms of θ there is no such normalization. In fact, the integral can be divergent, consider the example of a Pareto density with fixed scale $x_m = 1$, $f_\alpha(y) = \alpha/y^{\alpha+1}$ for $y \geq 1$ in which case we have for a single observation $y_1 = 1$ that

$$\int_{\alpha>0} f_\alpha(1) d\alpha = \infty.$$

(c) Let's practice performing MLE with a Poisson distribution, with a PMF given as: $f_\lambda(y) = \frac{\lambda^y e^{-\lambda}}{y!}$. Let $Y_1, Y_2, \ldots, Y_n$ be a set of independent and identically distributed random variables with Poisson distribution with parameter λ.

Find the joint distribution of $Y_1, Y_2, \ldots, Y_n$. Find the maximum likelihood estimate of λ as a function of observations $y_1, y_2, \ldots, y_n$.

Solution:

The joint probability mass function is the product of the probability mass functions of all n independent variables y_i,

$p_\theta(y_1, y_2, \ldots, y_n) = \prod_{i=1}^{n} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!}$.

The log likelihood will thus be $\ell(\lambda) = \sum_{i=1}^{n}(y_i \log(\lambda) - \lambda - \log(y_i!))$

We find the maximum by finding the derivative and setting it to 0:

$\ell'(\lambda) = (\sum_{i=1}^{n} \frac{y_i}{\lambda}) - n = 0$. Hence, the estimate should be $\hat{\lambda} = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{Y}$, which is the mean of the observations.

2 Linear Regression from MLE

In this problem, we will use maximum likelihood to motivate optimizing certain types of loss functions when performing linear regression.

To review, the goal of linear regression is to learn the parameters of a model $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ so that we can predict a label y that corresponds to input features $\mathbf{x}$, where $\mathbf{x} \in \mathbb{R}^d, y \in \mathbb{R}$. We are given a dataset of n input feature vectors and output labels $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$. (Note: we will ignore the bias term in this problem for simplicity.)

When performing linear regression on a dataset such as this one, we assume that the labels y_i are noisy: $y_i = \mathbf{w}^{*\top}\mathbf{x}_i + \varepsilon_i$, where $\mathbf{w}^*$ is the true linear model parameter we are trying to estimate.

There are many possible assumptions we could make about the noise ε_i, and in this problem, we will explore the implications of assuming noise sampled from certain probability distributions.

(a) **Assuming Gaussian noise variables $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, write the likelihood function $\mathcal{L}(\mathbf{w})$ of the dataset given a set of parameters $\mathbf{w}$.**

$$\mathcal{L}(\mathbf{w}) = p(\mathbf{x}_1, \dots, \mathbf{x}_n, y_1, \dots, y_n; \mathbf{w}) \tag{1}$$

The PDF of the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ is given as:

$$f(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(z-\mu)^2}{\sigma^2}\right\} \tag{2}$$

Hint: What is the probability distribution that gives $p(y_i|\mathbf{x}_i; \mathbf{w})$?

Hint: Remember that we can exclude all terms in our likelihood that don't depend on the parameter $\mathbf{w}$.

Solution:
Given a parameter $\mathbf{w}$, we first find the conditional distribution of $Y_i \mid X_i$ using the relationship $y_i = \mathbf{w}^\top \mathbf{x}_i + \varepsilon_i = \mathbf{w}^T \mathbf{x}_i + \mathcal{N}(0, \sigma^2)$.

Recall that given a normally distributed random variable $Z \sim \mathcal{N}(\mu, \sigma^2)$, adding a constant to this random variable gives a new random variable from the following distribution: $c + Z \sim \mathcal{N}(c + \mu, \sigma^2)$. (In our case, $c = \mathbf{w}^T \mathbf{x}_i$, and $Z = \varepsilon_i$.)

Therefore, $Y_i \mid X_i \sim \mathcal{N}(\mathbf{w}^T \mathbf{x}_i, \sigma^2)$, which means:

$$p(y_i|\mathbf{x}_i; \mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{\sigma^2}\right\} \tag{3}$$

Next, we compute the likelihood function, simplifying by using independence of the data points and the chain rule (splitting the joint probability into a conditional and a marginal probability):

$$\mathcal{L}(\mathbf{w}) = p(\mathbf{x}_1, \dots, \mathbf{x}_n, y_1, \dots, y_n; \mathbf{w}) \tag{4}$$

$$
\begin{aligned}
&= \prod_{i=1}^{n} p(\mathbf{x}_i, y_i; \mathbf{w}) && (5)\\
&= \prod_{i=1}^{n} p(y_i|\mathbf{x}_i; \mathbf{w}) p(\mathbf{x}_i) && (6)\\
&\propto \prod_{i=1}^{n} p(y_i|\mathbf{x}_i; \mathbf{w}) && (7)\\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{\sigma^2}\right\} && (8)
\end{aligned}
$$

The reason for excluding the $p(\mathbf{x}_i)$ term, which is the prior probability of the data, is that there is no dependence on the parameters of the model. Therefore, it will not affect the maximization.

(b) **Compute the log likelihood $\ell(\mathbf{w}) = \log \mathcal{L}(\mathbf{w})$ and simplify to matrix notation to show that the maximum likelihood objective is equivalent to:**

$$
\hat{\mathbf{w}}_{MLE} = \arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 \tag{9}
$$

$\mathbf{X}$ is an $n \times d$ matrix where each row represents a data point, and $\mathbf{y}$ is a n-dimensional vector of labels for all the data points.

Solution:

$$
\begin{aligned}
\ell(\mathbf{w}) &= \log \mathcal{L}(\mathbf{w}) && (10)\\
&= \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{\sigma^2}\right\} && (11)\\
&= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 && (12)\\
&= n \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 && (13)\\
& && (14)\\
\hat{\mathbf{w}}_{MLE} &= \arg\max_{\mathbf{w}} \; \ell(\mathbf{w}) && (15)\\
&= \arg\max_{\mathbf{w}} \; n \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 && (16)\\
&= \arg\max_{\mathbf{w}} \; -\sum_{i=1}^{n} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 && (17)\\
&= \arg\min_{\mathbf{w}} \; \sum_{i=1}^{n} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 = \arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 && (18)
\end{aligned}
$$

The steps above remove any terms that don't depend on the parameters, as well as scaling constants that don't affect the maximization. Flipping the sign of the objective also changes the problem from a maximization to a minimization.

(c) Now, we will assume a different noise distribution (still zero-mean and independently sampled). **Given noise variables sampled from a Laplace distribution $\varepsilon_i \overset{i.i.d.}{\sim}$ Laplace$(0, b)$, what is the likelihood function $\mathcal{L}(\mathbf{w})$?**

The PDF of the Laplace distribution Laplace(μ, b) is given as:

$$f(z) = \frac{1}{2b} \exp\left\{-\frac{|z-\mu|}{b}\right\} \tag{19}$$

Solution:
By similar logic as part (a), $Y_i \mid X_i \sim$ Laplace$(\mathbf{w}^T\mathbf{x}_i, b)$, which means:

$$p(y_i|\mathbf{x}_i; \mathbf{w}) = \frac{1}{2b} \exp\left\{-\frac{|y_i - \mathbf{w}^\top \mathbf{x}_i|}{b}\right\} \tag{20}$$

$$\mathcal{L}(\mathbf{w}) = \prod_{i=1}^{n} \frac{1}{2b} \exp\left\{-\frac{|y_i - \mathbf{w}^\top \mathbf{x}_i|}{b}\right\} \tag{21}$$

(d) **Compute the log likelihood $\ell(\mathbf{w})$ and simplify to matrix notation to show that the maximum likelihood objective is equivalent to:**

$$\hat{\mathbf{w}}_{MLE} = \arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_1 \tag{22}$$

Solution:

$$\begin{aligned}
\ell(\mathbf{w}) &= \log \mathcal{L}(\mathbf{w}) && (23)\\
&= \log \prod_{i=1}^{n} \frac{1}{2b} \exp\left\{-\frac{|y_i - \mathbf{w}^\top \mathbf{x}_i|}{b}\right\} && (24)\\
&= n \log \frac{1}{2b} - \frac{1}{b} \sum_{i=1}^{n} |y_i - \mathbf{w}^\top \mathbf{x}_i| && (25)\\
& && (26)\\
\hat{\mathbf{w}}_{MLE} &= \arg\max_{\mathbf{w}} \ \ell(\mathbf{w}) && (27)\\
&= \arg\max_{\mathbf{w}} n \log \frac{1}{2b} - \frac{1}{b} \sum_{i=1}^{n} |y_i - \mathbf{w}^\top \mathbf{x}_i| && (28)
\end{aligned}$$

$$= \arg\max_{\mathbf{w}} -\sum_{i=1}^{n} |y_i - \mathbf{w}^\top \mathbf{x}_i| \tag{29}$$

$$= \arg\min_{\mathbf{w}} \sum_{i=1}^{n} |y_i - \mathbf{w}^\top \mathbf{x}_i| = \arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_1 \tag{30}$$

(e) The ℓ-1 (least absolute deviations) objective is known to be more robust in the presence of outlier labels that fall far from the true distribution. **Why, from the viewpoint of maximum likelihood and probability densities, does it make sense that the ℓ-1 objective is less sensitive to outliers than the ℓ-2 (least squares) objective?**

Solution:
Compare the two conditional probabilities that make up each term of the likelihood function.

$$p(y_i|\mathbf{x}_i; \mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{\sigma^2}\right\} \tag{31}$$

$$p(y_i|\mathbf{x}_i; \mathbf{w}) = \frac{1}{2b} \exp\left\{-\frac{|y_i - \mathbf{w}^\top \mathbf{x}_i|}{b}\right\} \tag{32}$$

If $|y_i - \mathbf{w}^\top \mathbf{x}_i|$ is large (i.e. we are calculating this for an outlier), the quadratic error term in the Gaussian PDF blows up much faster than the absolute value in the Laplace PDF. Since this term is in an exponent with a negative sign in front, a huge negative exponent generates a very small likelihood (when multiplying a very small probability density). The ℓ-2 objective is much more sensitive to outliers, since they affect the likelihood probability more than ℓ-1 objective.

3 Maximum Likelihood Estimation for Reliability Testing

Suppose we are reliability testing n units taken randomly from a population of identical appliances. We want to estimate the mean failure time of the population. We assume the failure times come from an exponential distribution with parameter $\lambda > 0$, whose probability density function is $f(t) = \lambda e^{-\lambda t}$ on the domain $t \geq 0$.

(a) In an ideal (but impractical) scenario, we run the units until they all fail.

The failure times $T_1, T_2, \ldots, T_n$ for units $1, 2, \ldots, n$ are observed to be $t_1, t_2, \ldots, t_n$.

Formulate the likelihood function $\mathcal{L}(\lambda; t_1, \ldots, t_n)$ for our data. Then, find the maximum likelihood estimate $\hat{\lambda}$ for the distribution's parameter.

(Remember that it's equivalent, and usually easier, to optimize the log-likelihood.)

Solution:

$$\mathcal{L}(\lambda; t_1, \ldots, t_n) = \prod_{i=1}^{n} f(t_i) = \prod_{i=1}^{n} \lambda e^{-\lambda t_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} t_i}$$

$$\ln \mathcal{L}(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} t_i$$

$$\frac{\partial}{\partial \lambda} \ln \mathcal{L}(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} t_i = 0$$

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} t_i}.$$

(b) In a more realistic scenario, we run the units for a fixed time h. The failure time for $T_1, T_2, \ldots, T_r$ are observed to be $t_1, t_2, \ldots, t_r$, where $0 \leq r \leq n$. The remaining $n - r$ units survive the entire time h without failing. Let's find the maximum likelihood estimate $\hat{\lambda}$ for our model distribution parameters!

(a) **What is the probability that a unit will not fail during time h?**

Solution: $P(T > h) = 1 - P(T \leq h) = 1 - \int_{t=0}^{h} f(t)dt = 1 - \left[e^{-\lambda t}\right]_0^h = 1 - (1 - e^{-\lambda h}) = e^{-\lambda h}$.

(b) **Write the new likelihood function $\mathcal{L}(\lambda; h, n, r, t_1, \ldots, t_r)$ and optimize to find the MLE estimate $\hat{\lambda}$.**

Solution:

$$\begin{aligned}
\mathcal{L}(\lambda; n, h, r, t_1, \ldots, t_r) &= P(T_1 = t_1, \ldots, T_r = t_r, T_{i>r} > h; \lambda) \\
&= \left(\prod_{i=1}^{r} f(t_i)\right) P(t > h)^{n-r} \\
&= \left(\prod_{i=1}^{r} \lambda e^{-\lambda t_i}\right) \left(e^{-\lambda h}\right)^{n-r} \\
&= \lambda^r e^{-\lambda \sum_{i=1}^{r} t_i} e^{-\lambda(n-r)h}.
\end{aligned}$$

Note: The wording of the question specifies which units failed, therefore we shouldn't add the "n choose r" type of coefficient. Still, check in next question that even if you weren't told which were the units that failed when, the MLE solution would be the same!

(c) **Compare the two MLE estimates, and explain the difference with a physical interpretation.**

Solution:

$$\begin{aligned}
\ln \mathcal{L}(\lambda) &= r \ln \lambda - \lambda \sum_{i=1}^{r} t_i - \lambda(n-r)h \\
\frac{\partial}{\partial \lambda} \ln \mathcal{L}(\lambda) &= \frac{r}{\lambda} - \sum_{i=1}^{r} t_i - (n-r)h = 0 \\
\hat{\lambda} &= \frac{r}{\sum_{i=1}^{r} t_i + (n-r)h}.
\end{aligned}$$

We can interpret $\hat{\lambda}$ to be the number of observed failures divided by the sum of unit test times.