CS 189/289A Introduction to Machine Learning
Fall 2024 Jennifer Listgarten, Saeed Saremi

DIS1

1 Maximizing Likelihood & Minimizing Cost

Maximum Likelihood Estimation (MLE) is a method for estimating the parameters of a statistical model given observations.

Data Suppose we obtain n discrete *observations* belonging to $B := \{1, 2, 3, 4\}$. Our dataset looks something like the following.

$$\begin{aligned} r_1 &= 1 \\ r_2 &= 1 \\ r_3 &= 3 \\ &\vdots \\ r_n &= 1 \end{aligned}$$

Assumptions Suppose we aim to estimate the occurence probabilities of each class in B based on the observed data. We additionally assume that observations are independent and identically distributed (i.i.d.). In particular, this assumption implies that the order of the data does not matter.

Model Based on these assumptions, a natural model for our data is the multinomial distribution. In a multinomial distribution, the order of the data does not matter, and we can equivalently represent our dataset as $(y, c_y)_{y \in B}$, where c_y is the number of items of class y.

The probability mass function (PMF) of the multinomial distribution—this is, the probability in n trials of obtaining each class i x_i times—is

$$P(x_1, \ldots, x_k) = n! \prod_{i=1}^{k} \frac{p_i^{x_i}}{x_i!}.$$

(a) Derive an expression for the likelihood for this problem. What are the observations? What are the parameters? What parameters are we trying to estimate with MLE?

Solution: Since we are working in the discrete case, we simply have $\mathcal{L}(\theta; x) = P(x; \theta)$.

$$\mathcal{L}(p_1, p_2, p_3, p_4; c_1, c_2, c_3, c_4) = n! \prod_{y \in B} \frac{p_y^{c_y}}{c_y!}$$

The observations are the counts of each class c_1, c_2, c_3, c_4. The parameters are the probabilities of each class p_1, p_2, p_3, p_4, the total number of trials n and the total number of classes k. Out of these, we only aim to estimate the probabilities of each class p_1, p_2, p_3, p_4 (n and k are known).

(b) Typically, the log-likelihood $\ell(\theta) = \log L(\theta)$ is used instead of $L(\theta)$. Write down the expression for $\ell(\theta)$. Why might this be a good idea?

Solution:

$$\begin{aligned}\ell(p_1, p_2, p_3, p_4; c_1, c_2, c_3, c_4) &= \log\left(n! \prod_{y \in B} \frac{p_y^{c_y}}{c_y!}\right) \\ &= \log n! + \sum_{y \in B} c_y \log p_y - \sum_{y \in B} \log c_y!\end{aligned}$$

This program is concave in p_1, p_2, p_3, p_4 and has linear constraints (the sum of the probabilities is 1), which makes it easy to solve using convex optimization techniques.

(c) Another idea might be to minimize the cross-entropy based on raw observations, corresponding to the following program

$$\underset{\substack{p \in \mathbb{R}_+^4 \\ \|p\|_1 = 1}}{\operatorname{argmin}} - \sum_{i=1}^{n} \sum_{y \in B} \delta_{r_i y} \log p_y$$

where p is the vector of probabilities per class $\begin{bmatrix} p_1 & p_2 & p_3 & p_4 \end{bmatrix}^\top$, and $\delta_{r_i y}$ is the Kronecker delta that outputs 1 if $r_i = y$ and 0 otherwise.

Show that this program is equivalent to the MLE program.

Solution: MLE and maximum entropy actually provide the same estimates for this problem.

$$\begin{aligned}\underset{\substack{p \in \mathbb{R}_+^4 \\ \|p\|_1 = 1}}{\operatorname{argmax}} \mathcal{L}(p; c_1, c_2, c_3, c_4) &= \underset{\substack{p \in \mathbb{R}_+^4 \\ \|p\|_1 = 1}}{\operatorname{argmax}} n! \prod_{y \in B} \frac{p_y^{c_y}}{c_y!} \\ &= \underset{\substack{p \in \mathbb{R}_+^4 \\ \|p\|_1 = 1}}{\operatorname{argmax}} \log \prod_{y \in B} \frac{p_y^{c_y}}{c_y!} \\ &= \underset{\substack{p \in \mathbb{R}_+^4 \\ \|p\|_1 = 1}}{\operatorname{argmax}} \sum_{y \in B} \log \frac{p_y^{c_y}}{c_y!} \\ &= \underset{\substack{p \in \mathbb{R}_+^4 \\ \|p\|_1 = 1}}{\operatorname{argmax}} \sum_{y \in B} c_y \log p_y - \sum_{y \in B} \log\left(c_y!\right)\end{aligned}$$

Note that the $\sum_{y \in B} \log\left(c_y!\right)$ term is a constant with respect to p, so it does not affect the optimization problem.

$$\begin{aligned}\underset{\substack{p \in \mathbb{R}_+^4 \\ \|p\|_1 = 1}}{\operatorname{argmax}} \sum_{y \in B} c_y \log p_y - \sum_{y \in B} \log\left(c_y!\right) &= \underset{\substack{p \in \mathbb{R}_+^4 \\ \|p\|_1 = 1}}{\operatorname{argmin}} - \sum_{y \in B} c_y \log p_y \\ &= \underset{\substack{p \in \mathbb{R}_+^4 \\ \|p\|_1 = 1}}{\operatorname{argmin}} - \sum_{i=1}^{n} \sum_{y \in B} \delta_{r_i y} \log p_y\end{aligned}$$

■

2 Independence and Multivariate Gaussians

As described in lecture, a covariance matrix $\Sigma \in \mathbb{R}^{N\times N}$ for a random variable $X \in \mathbb{R}^N$ with the following values, where $\text{cov}(X_i, X_j) = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]$ is the covariance between the i-th and j-th elements of the random vector X:

$$\Sigma = \begin{bmatrix} \text{cov}(X_1, X_1) & \dots & \text{cov}(X_1, X_n) \\ \dots & & \dots \\ \text{cov}(X_n, X_1) & \dots & \text{cov}(X_n, X_n) \end{bmatrix}. \tag{1}$$

Recall that the density of an N dimensional Multivariate Gaussian Distribution $\mathcal{N}(\mu, \Sigma)$ is defined as follows when Σ is positive definite:

$$f(x) = \frac{1}{\sqrt{(2\pi)^N |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}. \tag{2}$$

Here, $|\Sigma|$ denotes the determinant of the matrix Σ.

(a) Consider the random variables X and Y in $\mathbb{R}$ with the following conditions.

(i) X and Y can take values $\{-1, 0, 1\}$.

(ii) When X is 0, Y takes values 1 and -1 with equal probability ($\frac{1}{2}$). When Y is 0, X takes values 1 and -1 with equal probability ($\frac{1}{2}$).

(iii) Either X is 0 with probability ($\frac{1}{2}$), or Y is 0 with probability ($\frac{1}{2}$).

Are X and Y uncorrelated? Are X and Y independent? Prove your assertions. *Hint:* Write down the joint probability of (X, Y) for each possible pair of values they can take.

Solution: Essentially, there are 4 possible pairs of points that (X, Y) can be, all with equal probability ($\frac{1}{4}$): $\{(0, 1), (0, -1), (1, 0), (-1, 0)\}$, If graphed onto the Cartesian plane, these points will form "crosshairs".

To show that X and Y are uncorrelated, we need to prove:

$$\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = 0,$$

or equivalently, that $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$.

Since X, Y can be simultaneously nonzero, $\mathbb{E}[XY] = 0$.

In parallel,

$$\mathbb{E}[X] = \mathbb{E}[Y] = \frac{1}{2} * 0 + \frac{1}{2} * \left(\frac{1}{2} + \frac{-1}{2}\right) = 0$$

We have shown that X and Y are uncorrelated.

X and Y would be independent if

$$P(X|Y) = P(X)$$

Unfortunately, this is not the case. $P(X = 0) = \frac{1}{2}$, but $P(X = 0|Y = 1) = 1$. Thus, X and Y are not independent.

(b) For $X = [X_1, \cdots, X_n]^\top \sim \mathcal{N}(\mu, \Sigma)$, **verify that if X_i, X_j are independent (for all $i \neq j$), then Σ must be diagonal, i.e., X_i, X_j are uncorrelated.**

Solution: Recall that if random variables Z, W are independent, we have $\mathbb{E}[ZY] = \mathbb{E}[Z]\,\mathbb{E}[Y]$. Since the covariance $\mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)] = \mathbb{E}[X_i - \mu_i]\,\mathbb{E}[X_j - \mu_j] = 0 \cdot 0$ is 0, it follows that the pair of variables X_i, X_j are uncorrelated.

(c) Let $N = 2$, $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, and $\Sigma = \begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix}$. Suppose $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma)$. **Show that X_1, X_2 are independent if $\beta = 0$.** Recall that two continuous random variables W, Y with joint density $f_{W,Y}$ and marginal densities f_W, f_Y are independent if $f_{W,Y}(w, y) = f_W(w) f_Y(y)$.

Solution: Recall that the marginal density of two jointly Gaussian random variables is also Gaussian. In particular, we have that $X_1 \sim \mathcal{N}(\mu_1, \alpha)$ and $X_2 \sim \mathcal{N}(\mu_2, \gamma)$. Let's denote the marginal densities as $f_{X_1}(\cdot)$ and $f_{X_2}(\cdot)$.

Since $\beta = 0$, we may compute the inverse $\Sigma^{-1} = \begin{pmatrix} \alpha^{-1} & 0 \\ 0 & \gamma^{-1} \end{pmatrix}$.

Let's write out the joint density of X_1, X_2:

$$\begin{aligned} f_{X_1,X_1}(x_1, x_2) &= \frac{1}{\sqrt{(2\pi)^N |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)} \\ &= \frac{1}{\sqrt{(2\pi)^2 \alpha\gamma}} e^{-\frac{1}{2}(\alpha^{-1}(x_1-\mu_1)^2 + \gamma^{-1}(x_2-\mu_2)^2)} \\ &= \frac{1}{\sqrt{(2\pi)\alpha}} e^{-\frac{(x_1-\mu_1)^2}{2\alpha}} \cdot \frac{1}{\sqrt{(2\pi)\gamma}} e^{-\frac{(x_2-\mu_2)^2}{2\gamma}} \\ &= f_{X_1}(x_1) \cdot f_{X_2}(x_2) \end{aligned}$$

This proves that X_1, X_2 are independent if $\beta = 0$. Note that we don't need to verify that $f_{X_1}(x_1)$ and $f_{X_2}(x_2)$ are properly normalized (i.e. integrate to 1), since we can always shift around constant factors to ensure that this is the case.

(d) Consider a data point x drawn from an N-dimensional zero mean Multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$, as shown above. Assume that Σ^{-1} exists. **Prove that there exists a matrix $A \in \mathbb{R}^{N \times N}$ such that $x^\top \Sigma^{-1} x = \|Ax\|_2^2$ for all vectors x. What is the matrix A?**

Solution: Use the Spectral Theorem to decompose Σ into a product involving the following: an orthonormal matrix Q of orthonormal eigenvectors $\mathbf{v_i}$ $\forall i \in [1...N]$ and a diagonal matrix D with eigenvalues λ_i $\forall i \in [1...N]$ along the diagonal. Note that all the eigenvalues are strictly

positive since Σ is positive definite (it is a covariance matrix and Σ^{-1} exists). Hence, we may write

$$\Sigma = QDQ^\top,$$

and, therefore,

$$\Sigma^{-1} = (QDQ^\top)^{-1} = (Q^\top)^{-1}D^{-1}Q^{-1} = QD^{-1}Q^\top.$$

This is because orthonormal matrices satisfy $Q^{-1} = Q^\top$.

Note that if the matrix D has values λ_i along its diagonal, then D^{-1} has values $\frac{1}{\lambda_i}$ along its diagonal. Once again, since Σ was positive definite, the reciprocal $\frac{1}{\lambda_i}$ exists (each $\lambda_i > 0$).

Now, we can decompose D^{-1} into its square-root by defining S as a diagonal matrix with diagonal values $\frac{1}{\sqrt{\lambda_i}}$. You can quickly verify that $SS = D^{-1}$ and that $S^\top = S$. Thus, we have,

$$\Sigma^{-1} = QD^{-1}Q^\top = QSSQ^\top = QSS^\top Q^\top \tag{3}$$

$$\Sigma^{-1} = A^\top A, \tag{4}$$

where we let $A = (QS)^\top$. Therefore,

$$x^\top \Sigma^{-1} x = x^\top A^\top A x = (Ax)^\top (Ax) = \|Ax\|_2^2. \tag{5}$$

Note that A is not necessarily unique, however, since, if $A^\top A = \Sigma^{-1}$, then $(QA)^\top QA = A^T Q^T QA = A^T(I)A = A^T A = \Sigma^{-1}$ as well for any orthonormal Q.

3 Least Squares (using vector calculus)

In ordinary least-squares linear regression, we typically have $n > d$ so that there is no $\mathbf{w}$ such that $\mathbf{Xw} = \mathbf{y}$ (these are typically overdetermined systems — too many equations given the number of unknowns). Hence, we need to find an approximate solution to this problem. The residual vector will be $\mathbf{r} = \mathbf{Xw} - \mathbf{y}$ and we want to make it as small as possible. The most common case is to measure the residual error with the standard Euclidean ℓ^2-norm. So the problem becomes:

$$\min_{\mathbf{w}} \|\mathbf{Xw} - \mathbf{y}\|_2^2,$$

where $\mathbf{X} \in \mathbb{R}^{n \times d}, \mathbf{w} \in \mathbb{R}^d, \mathbf{y} \in \mathbb{R}^n$.

Assume that $\mathbf{X}$ is full rank.

(a) How do we know that $\mathbf{X}^\top\mathbf{X}$ is invertible?

Solution: Matrix $\mathbf{X}$ is said to be full rank if n ≥ d and its columns are not linear combinations of each other. In this case, $\mathbf{X}^\top\mathbf{X}$ will be positive definite and therefore invertible. If $\mathbf{X}$ is not full rank, at least one of the columns will be a linear combination of the other columns. In this case, the rank of $\mathbf{X}$ will be less than n and $\mathbf{X}^\top\mathbf{X}$ will not be invertible.

In this question, we know that $\mathbf{X}$ has full rank, so if we can show that the rank of $\mathbf{X}$ is equivalent to the rank of $\mathbf{X}^\top\mathbf{X}$, then $\mathbf{X}^\top\mathbf{X}$ has full rank and is therefore invertible. Let us show the ranks are equivalent using nullspaces. Suppose $\mathbf{v}$ is in the nullspace of $\mathbf{X}^\top\mathbf{X}$ meaning $\mathbf{X}^\top\mathbf{Xv} = \mathbf{0}$:

$$\begin{aligned}
\mathbf{X}^\top\mathbf{Xv} &= \mathbf{0} \\
\mathbf{v}^\top\mathbf{X}^\top\mathbf{Xv} &= 0 \\
(\mathbf{Xv})^\top(\mathbf{Xv}) &= 0 \\
\|\mathbf{Xv}\|_2^2 &= 0 \\
\mathbf{Xv} &= 0 \qquad \text{Because the only vector whose length is 0 is the } \mathbf{0} \text{ vector.}
\end{aligned}$$

From this we can see that any $\mathbf{v}$ which is in nullspace of $\mathbf{X}^\top\mathbf{X}$ also needs to be in the nullspace of $\mathbf{X}$. Since $\mathbf{X}$ and $\mathbf{X}^\top\mathbf{X}$ have the same null space, then $\mathbf{X}^\top\mathbf{X}$ should also be full rank and therefore invertible.

(b) Derive using vector calculus an expression for an optimal estimate for $\mathbf{w}$ for this problem.

Solution: The work flow is as follows: We first find a critical point by setting the gradient to 0, then show that it is unique under the conditions in the question and finally that it is in fact a minimizer.

Let us first find critical points $\mathbf{w}_{OLS}$ such that the gradient is zero, i.e $\nabla_w \|\mathbf{Xw}_{OLS} - \mathbf{y}\|_2^2\big|_{\mathbf{w}=\mathbf{w}_{OLS}} = 0$. In order to take the gradient, we expand the ℓ^2-norm. First, note the following:

$$\nabla_w(\mathbf{w}^\top\mathbf{Bw}) = (\mathbf{B} + \mathbf{B}^\top)\mathbf{w}$$

$$\nabla_w(\mathbf{w}^\top \mathbf{b}) = \mathbf{b}$$

We start by expanding the ℓ^2-norm:

$$\begin{aligned}
&\nabla_w(\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y}) \\
&= \nabla_w((\mathbf{X}\mathbf{w})^T(\mathbf{X}\mathbf{w}) - (\mathbf{X}\mathbf{w})^T(\mathbf{y}) - \mathbf{y}^T(\mathbf{X}\mathbf{w}) + \mathbf{y}^T\mathbf{y}) && \text{Combine middle terms, identical scalars.} \\
&= \nabla_w(\mathbf{w}^T\mathbf{X}^\top\mathbf{X}\mathbf{w} - 2\mathbf{w}^T\mathbf{X}^\top\mathbf{y} + \mathbf{y}^\top\mathbf{y}) && \text{Apply two derivative rules above} \\
&= (\mathbf{X}^\top\mathbf{X} + \mathbf{X}^\top\mathbf{X})\mathbf{w} - 2\mathbf{X}^\top\mathbf{y} \\
&= 2\mathbf{X}^\top(\mathbf{X}\mathbf{w} - \mathbf{y})
\end{aligned}$$

Having computed the gradient, we now require it to vanish at the critical point $\mathbf{w} = \mathbf{w}_{OLS}$

$$\begin{aligned}
\nabla_w \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2\big|_{\mathbf{w}=\mathbf{w}_{OLS}} &= 2\mathbf{X}^T(\mathbf{X}\mathbf{w}_{OLS} - \mathbf{y}) \\
&= 2\mathbf{X}^\top\mathbf{X}\mathbf{w}_{OLS} - 2\mathbf{X}^\top\mathbf{y} = 0 \\
\implies \mathbf{X}^\top\mathbf{X}\mathbf{w}_{OLS} &= \mathbf{X}^\top\mathbf{y}
\end{aligned}$$

Because X is full rank, $\mathbf{X}^T\mathbf{X}$ is invertible (see question (b)) and thus there is only one vector which satisfies the last equation which reads: $\mathbf{w}_{OLS} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$. Therefore, there is only one unique critical point.

Furthermore, observe that the least square is twice differentiable, and that its Hessian corresponds to

$$\begin{aligned}
\nabla_w^2(\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y}) &= \nabla_w\nabla_w(\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y}) && (6) \\
&= \nabla_w(2\mathbf{X}^\top(\mathbf{X}\mathbf{w} - \mathbf{y})) && (7) \\
&= 2\mathbf{X}^T\mathbf{X}. && (8)
\end{aligned}$$

This matrix is positive definite, implying that the objective function is convex. We conclude from this observation that $\mathbf{w}_{OLS}$ is the global minimum.

(c) What should we do if $\mathbf{X}$ is not full rank?

Solution: (Basic idea) If $\mathbf{X} \in \mathbf{R}^{n\times d}$ is not full rank, there is no unique answer. As we will see later, this is not an issue in ridge regression where we add a penalization to the loss function (thus change the loss function) which forces a unique solution. Another possibility is to use the solution that minimizes the norm of $\mathbf{w}$ (in later lectures we will see why that might be a good thing to do).

The minimum norm solution can be found by using the pseudo-inverse of $\mathbf{X}^\top\mathbf{X}$. The pseudo-inverse of an arbitrary matrix $\mathbf{X}$ is denoted as $\mathbf{X}^\dagger$. More intuitively, $\mathbf{X}^\dagger$ behaves most similarly to the inverse: it is the matrix that, when multiplied by $\mathbf{X}$, minimizes distance to the identity. $\mathbf{X}^\dagger = \text{argmin}_{W\in\mathbf{R}^{n\times d}}\|\mathbf{X}\mathbf{W} - \mathbf{I}_m\|_F$.