CS 189 Introduction to Machine Learning
Spring 2022 Marvin Zhang

DIS2

1 MLE vs. MAP

Let D denote the observed data and θ the parameter. While MLE only maximizes a likelihood distribution $p(D|\theta)$, MAP takes a more Bayesian approach. MAP assumes that the parameter θ *is also a random variable and has its own distribution*. Recall that using Bayes' rule, the posterior distribution can be seen as the product of likelihood and prior:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \propto \underbrace{p(D|\theta)}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}}$$

Suppose that the data consists of n i.i.d. observations $D = \{x_1, \ldots, x_n\}$. MAP tries to infer the parameter by maximizing the posterior distribution:

$$\begin{aligned}
\hat{\theta}_{\text{MAP}} &= \arg\max_{\theta} \; p(\theta|D) \\
&= \arg\max_{\theta} \; p(D|\theta)p(\theta) \\
&= \arg\max_{\theta} \left[\prod_{i=1}^{n} p(x_i|\theta)\right] p(\theta) \\
&= \arg\max_{\theta} \left(\sum_{i=1}^{n} \log p(x_i|\theta)\right) + \log p(\theta)
\end{aligned}$$

Note that since both of these methods are point estimates (they yield a value rather than a distribution), neither of them are completely Bayesian. A faithful Bayesian would use a model that yields a posterior distribution over all possible values of θ, but this is often intractable or very computationally expensive.

Now suppose we have a coin with unknown bias θ. We are trying to find the bias of the coin by maximizing the underlying distribution. You tossed the coin $n = 10$ times and 3 of the tosses came as heads.

(a) **What is the MLE of the bias of the coin $\hat{\theta}_{\text{MLE}}$?**

Solution:

$$p(x|\theta) \propto \theta^x(1-\theta)^{(n-x)} = \theta^3(1-\theta)^7.$$

Taking the logarithm for easier computation, we have

$$\log p(x|\theta) = 3\log\theta + 7\log(1-\theta) + C.$$

This is a concave function and thus the maximum is achieved by setting the derivative w.r.t. θ to 0:

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(x|\theta) = \frac{3}{\theta} - \frac{7}{1-\theta} = 0.$$

Therefore,

$$\hat{\theta}_{\text{MLE}} = 0.3.$$

(b) Suppose we know that the bias of the coin is distributed according to $\theta \sim N(0.8, 0.09)$, i.e., we are rather sure that the bias should be around 0.8.[1]

What is the MAP estimate of the coin bias $\hat{\theta}_{\text{MAP}}$? You can leave your result as a polynomial equation on θ.

Solution: Now take into account the prior distribution:

$$\begin{aligned} p(\theta|x) &\propto p(x|\theta)p(\theta) \\ &\propto \theta^x(1-\theta)^{n-x} \exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2}\right] \\ &= \theta^3(1-\theta)^7 \exp\left[-\frac{(\theta-0.8)^2}{2\times 0.09}\right]. \end{aligned}$$

Taking the logarithm,

$$\log p(\theta|x) = 3\log\theta + 7\ln(1-\theta) - \frac{(\theta-0.8)^2}{2\times 0.09} + C.$$

Taking the derivative w.r.t. θ,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log p(\theta|x) = \frac{3}{\theta} - \frac{7}{1-\theta} - \frac{\theta-0.8}{0.09} = 0.$$

Solving the equation yields

$$\hat{\theta}_{\text{MAP}} \approx 0.406.$$

$\hat{\theta}$ is now larger because we are assuming a larger prior.

(c) What if our prior is $\theta \sim N(0.5, 0.09)$ or $N(0.8, 1)$ instead?

How does the difference between the new MAP estimates and MLE estimate change and why?

Solution: The above equation would instead be

$$\frac{3}{\theta} - \frac{7}{1-\theta} - \frac{\theta-0.5}{0.09} = 0$$

for $N(0.5, 0.09)$ and

$$\frac{3}{\theta} - \frac{7}{1-\theta} - (\theta-0.8) = 0$$

[1] This is a somewhat strange choice of prior, since we know that $0 \le \theta \le 1$. However, we will stick with this example for illustrative purposes.

for $N(0.8, 1)$. $\hat{\theta}_{\text{MAP}} \approx 0.340$ for $N(0.5, 0.3)$ and $\hat{\theta}_{\text{MAP}} \approx 0.31$ for $N(0.8, 1)$. For $N(0.5, 0.3)$, the prior is less distant from the experiment result; for $N(0.8, 1)$, the prior is weaker due to a larger variance. Therefore, the difference between the two models will decrease.

(d) **What if our prior is that θ is uniformly distributed in the range $(0, 1)$?**

Solution: The MLE and MAP estimate will be the same since the prior term $p(\theta)$ is uniform and can be canceled out. From a Bayesian perspective, MLE can, in certain cases, be seen as a special case of MAP estimation with a uniform prior.

2 Probabilistic Interpretation of Lasso

Let's start with the probabilistic interpretation of least squares. We're given labels $y \in \mathbb{R}$, data $\mathbf{x} \in \mathbb{R}^d$, and Gaussian noise $z \sim \mathcal{N}(0, \sigma^2)$, where $y = \mathbf{w}^T\mathbf{x} + z$. Recall from lecture and the previous discussion that this results in a probabilistic linear model given by:

$$p(y|\mathbf{x}; \mathbf{w}) \sim \mathcal{N}(\mathbf{w}^\mathbf{T}\mathbf{x}, \sigma^2)$$

However, maximum likelihood estimates (MLE) can overfit to the training data (analagous to how fitting a very high dimensional polynomial to data leads to large coefficients and extreme behavior at unseen points). To ameliorate this issue, we can assume a zero-mean Laplace prior on each component of the parameter $w_j \sim \text{Laplace}(0, t)$:

$$p(w_j) = \frac{1}{2t}\exp\left\{-\frac{1}{t}|w_j|\right\}$$

$$p(\mathbf{w}) = \prod_{j=1}^{d} p(w_j) = (\frac{1}{2t})^d \cdot \exp\left\{-\frac{1}{t}\sum_{j=1}^{d}|w_j|\right\}$$

Assume that t is a known constant. Here, we will see that this modification results in a new objective called Lasso regression.

(a) Recall that the MLE objective finds the parameters that maximize the likelihood of the data,

$$\begin{aligned}\mathbf{w}^* &= \arg\max_{\mathbf{w}} \; L(\mathbf{w}) \\ &= \arg\max_{\mathbf{w}} \; p(Y_1, \ldots, Y_n, |\mathbf{w}, \mathbf{X_1}, \ldots, \mathbf{X_n}, \sigma^2) \\ &= \arg\max_{\mathbf{w}} \; \prod_{i=1}^{n} p(Y_i|\mathbf{X_i}, \mathbf{w}, \sigma^2).\end{aligned}$$

When working in a Bayesian framework, we instead focus on the posterior probability of the parameters (the unknown quantity) conditioned the data (the evidence):

$$\text{Posterior} = p(\text{unknowns} \mid \text{evidence}) = p(\mathbf{w}|Y_1, \ldots, Y_n, \mathbf{X_1}, \ldots, \mathbf{X_n}, \sigma^2)$$

Derive the MAP objective as a function of the log-likelihood $\ell(\mathbf{w})$ and the prior $p(\mathbf{w})$.

Solution:

$$\begin{aligned}\hat{\mathbf{w}}_{\text{MAP}} &= \arg\max_{\mathbf{w}} \; \frac{P(\mathbf{w}, Y_1, \ldots, Y_n|\mathbf{X_1}, \ldots, \mathbf{X_n}, \sigma^2)}{P(Y_1, \ldots, Y_n|\mathbf{X_1}, \ldots, \mathbf{X_n}, \sigma^2)} \\ &= \arg\max_{\mathbf{w}} \; \frac{P(Y_1, \ldots, Y_n, |\mathbf{w}, \mathbf{X_1}, \ldots, \mathbf{X_n}, \sigma^2)P(\mathbf{w})}{P(Y_1, \ldots, Y_n|\mathbf{X_1}, \ldots, \mathbf{X_n}, \sigma^2)} \\ &= \arg\max_{\mathbf{w}} \; \frac{L(\mathbf{w})P(\mathbf{w})}{P(Y_1, \ldots, Y_n)}\end{aligned}$$

$$= \arg\max_{\mathbf{w}} L(\mathbf{w})P(\mathbf{w}) \qquad \text{since } P(Y_1, \ldots, Y_n | \mathbf{X_1}, \ldots, \mathbf{X_n}, \sigma^2) \text{ does not depend on } \mathbf{w}.$$
$$= \arg\max_{\mathbf{w}} \ell(\mathbf{w}) + \log P(\mathbf{w})$$

We call $\mathbf{w}^*$ the Maximum a posteriori (MAP) estimate.

(b) **Fill in the terms of the MAP objective you derived, assuming Gaussian noise and a Laplace prior on the parameter.**

Solution:

$$P(w|\mathbf{X_i}, Y_i) \propto (\prod_{i=1}^{n} \mathcal{N}(Y_i|\mathbf{w^T X_i}, \sigma^2)) \cdot P(\mathbf{w}) = (\prod_{i=1}^{n} \mathcal{N}(Y_i|\mathbf{w^T X_i}, \sigma^2)) \cdot \prod_{j=1}^{D} P(w_j)$$

Taking the log of the above expression, we now have:

$$\begin{aligned} l(\mathbf{w}) &= \sum_{i=1}^{n} log\mathcal{N}(Y_i|\mathbf{w^T X_i}, \sigma^2) + \sum_{j=1}^{D} logP(w_j) \\ &= \sum_{i=1}^{n} log(\frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(Y_i - \mathbf{w^T X_i})^2}{2\sigma^2})) + \sum_{j=1}^{D} log(\frac{1}{2t} exp(\frac{-|w_j|}{t})) \\ &= -\sum_{i=1}^{n} \frac{(Y_i - \mathbf{w^T X_i})^2}{2\sigma^2} + \frac{-\sum_{j=1}^{D} |w_j|}{t} + nlog(\frac{1}{\sqrt{2\pi}\sigma}) + Dlog(\frac{1}{2t}) \end{aligned}$$

After dropping constants that don't depend on $\mathbf{w}$ and converting sums to their respective norms, we get:

$$\hat{\mathbf{w}}_{\text{MAP}} = \arg\max_{\mathbf{w}} -\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 - \frac{1}{t}\|\mathbf{w}\|_1$$

(c) **Using your answer from the previous part, show that maximizing the MAP objective is equivalent to minimizing the following:**

$$J(\mathbf{w}) = \sum_{i=1}^{n} (Y_i - \mathbf{w^T X_i})^2 + \lambda \|\mathbf{w}\|_1$$

What is the constant λ in terms of given quantities?

Solution:

$$\begin{aligned} \hat{\mathbf{w}}_{\text{MAP}} &= \arg\max_{\mathbf{w}} -\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 - \frac{1}{t}\|\mathbf{w}\|_1 \\ &= \arg\min_{\mathbf{w}} \frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \frac{1}{t}\|\mathbf{w}\|_1 \qquad \text{since } \arg\max_x f(x) = \arg\min_x -f(x) \\ &= \arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \underbrace{\frac{2\sigma^2}{t}}_{\lambda} \|\mathbf{w}\|_1 \qquad \text{since } 2\sigma^2 > 0 \end{aligned}$$

3 Independence and Multivariate Gaussians

To review, 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:

$$\Sigma = \begin{bmatrix} \text{cov}(X_1, X_1) & \dots & \text{cov}(X_1, X_n) \\ \dots & & \dots \\ \text{cov}(X_n, X_1) & \dots & \text{cov}(X_n, X_n) \end{bmatrix} = \mathbb{E}[(X-\mu)(X-\mu)^\top]\,. \tag{1}$$

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

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

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

(a) 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, that is, X_i, X_j are uncorrelated.** **Solution:** Recall that if random variables Z, W are independent, we have $\mathbb{E}[ZY] = \mathbb{E}[Z]\mathbb{E}[Y]$. Thus we have 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, i.e. the X_i, X_j are uncorrelated.

(b) 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 $\Sigma^{-1} = \begin{pmatrix} \alpha^{-1} & 0 \\ 0 & \gamma^{-1} \end{pmatrix}$.

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

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

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.

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

Solution: Use the Spectral Decomposition Theorem to convert Σ into the following: U is a unitary matrix of orthonormal eigenvectors $\mathbf{e_i}$ $\forall i \in [0...N]$ and D is a diagonal matrix with eigenvalues λ_i $\forall i \in [0...N]$ located at indices corresponding to eigenvectors in U. Note that all the eigenvalues are greater 0 since Σ is positive semidefine (it is a covariance matrix). Indeed, for any $v \in \mathbb{R}^n$, $v^\top \Sigma v = \mathbb{E}[v^\top (XX^\top) v] = \mathbb{E}[\|Xv\|_2^2] \geq 0$ (but this was not necessary to show to receive full credit). In fact, since Σ is invertible, all eigenvalues of Σ are strictly positive, Hence, we may write

$$\Sigma = UDU^\top,$$

and therefore

$$\Sigma^{-1} = (UDU^\top)^{-1} = (U^\top)^{-1} D^{-1} U^{-1} = UD^{-1}U^\top.$$

This is because a unitary matrix U is such that $U^{-1} = U^\top$. Note that if the diagonal matrix D has values $d_{i,i}$ $\forall i$, then D^{-1} has value $\frac{1}{d_{i,i}}$ $\forall i$. Once again, since Σ was positive definite, the value $\frac{1}{d_{i,i}}$ exists.

Now, we decompose D^{-1} into its square-root by defining Q as a diagonal matrix with diagonal values $\frac{1}{\sqrt{d_{i,i}}}$. Verify that $QQ = D^{-1}$ and that $Q^\top = Q$. Thus, we have:

$$\Sigma^{-1} = UD^{-1}U^\top = UQQU^\top = UQQ^\top U^\top \tag{3}$$

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

where we've defined $A = (UQ)^\top$. 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, since if $A^\top A = \Sigma^{-1}$, then also $(QA)^\top QA = \Sigma^{-1}$ for any orthogonal Q.

(d) Let's constrain x to be on the unit sphere. In other words, the ℓ_2 norm (or magnitude) of vector x is 1 ($\|x\|_2 = 1$). In this case, **what are the maximum and minimum values of $\|Ax\|_2^2$? In other words, $\max_{x:\|x\|_2=1} \|Ax\|_2^2$ and $\min_{x:\|x\|_2=1} \|Ax\|_2^2$?**

Solution:

$x^\top \Sigma^{-1} x$ is a scalar written in vector quadratic form. It looks like an incomprehensible value, but when we convert it to $\|Ax\|_2^2$, we see that in reality its just the squared L2 norm of Ax, which measures the squared distance from the data vector x from the mean (in this case 0). Note that we can change the mean to be any arbitrary value without loss of generality.

Recall from Part B our decomposition for Σ^{-1}, which was as follows where U is a unitary matrix, D is a diagonal matrix.

$$\Sigma^{-1} = UD^{-1}U^\top = A^\top A \tag{6}$$

Note that $\|x\|_2 = 1$ and $\|Ux\|_2 = 1$ since unitary matrices are orthonormal and preserve magnitude. Define $q = Ux$, we have

$$\|Ax\|_2^2 = x^\top A^\top Ax = x^\top UD^{-1}U^\top x = q^\top D^{-1}q \tag{7}$$

We can choose our x such that q will be any Euclidean Basis Vector $\mathbf{e_i}$ such that the ith element is 1 and all other elements are 0. Therefore, the maximum value that $\|Ax\|_2^2$ is $\frac{1}{\lambda_i}$, where λ_i is the minimum eigenvalue of Σ. The minimum value that $\|Ax\|_2^2$ is $\frac{1}{\lambda_j}$, where λ_j is the maximum eigenvalue of Σ.

(e) If we had $X_i \perp\!\!\!\perp X_j \ \forall i, j$ ($\perp\!\!\!\perp$ denotes independence), **what is the intuitive meaning for the maximum and minimum values of $\|Ax\|_2^2$?** Suppose you wanted to choose an x on the unit sphere to maximize the density function $f(x)$ in Eq (2); **what x should you choose?**

Solution:

As we showed in a previous part, if we have $X_i \perp\!\!\!\perp X_j \ \forall i, j$, then $cov(X_i, X_j) = 0 \ \forall i, j$, meaning that off diagonal terms for Σ are 0. Thus, we can find Σ^{-1} directly, as follows.

$$\Sigma_{i,j}^{-1} = \begin{cases} \frac{1}{\sigma_i^2} \text{ if } i = j \\ 0 \text{ else} \end{cases}$$

Therefore, if we have $X_i \perp\!\!\!\perp X_j \ \forall i, j$, the maximum value that $\|Ax\|_2^2$ is $\frac{1}{\sigma_i^2}$, where σ_i^2 is the minimum variance. The minimum value of $\|Ax\|_2^2$ is $\frac{1}{\sigma_j^2}$, where σ_j^2 is the maximum variance.

To maximize $f(x)$, we want the superscript above the exponent to be minimal since there is a negative sign. Thus, for $\|Ax\|_2^2$ to be minimal, we want to choose x to be the unit eigenvector corresponding to the maximal eigenvalue λ_j (i.e. maximum variance σ_j^2).