CS 189/289A Introduction to Machine Learning
Fall 2022 Jennifer Listgarten, Jitendra Malik

DIS2

1 MLE vs. MAP

Let D denote the observed data and θ the parameter. Whereas MLE only assumes and tries to maximize 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:

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

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 oftentimes intractable or very computationally expensive.

Now suppose we have a coin with unknown bias θ. We will estimate the bias of the coin using MLE and MAP. 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 θ?

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 θ? You can leave your result as an equation of the form $\frac{a}{\theta} - \frac{b}{1-\theta} - \frac{\theta-c}{d}$.

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,

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

Taking the derivative w.r.t. θ,

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

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 $N(0.8, 1)$? Write out the new equations using your previous answer, but you do not need to solve for the exact numeric value. How does the difference between MAP and MLE 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

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

for $N(0.8, 1)$. $\hat{\theta}_{\text{MAP}} \approx 0.340$ for $N(0.5, 0.09)$ and $\hat{\theta}_{\text{MAP}} \approx 0.31$ for $N(0.8, 1)$. For $N(0.5, 0.09)$, 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.

[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.

(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 Tikhonov Regularization

As defined in the homework, Tikhonov regularized regression is a generalization of ridge regression specified by the optimization problem

$$\arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{\Gamma}\mathbf{w}\|_2^2,$$

For some full rank matrix $\mathbf{\Gamma} \in \mathbb{R}^{d\times d}$.

In this problem, we look at Tikhonov regularization from a probabilistic standpoint and how it relates to the MAP estimator for a certain choice of prior on the parameters $\mathbf{w}$.

Let $\mathbf{x} \in \mathbb{R}^d$ be a d-dimensional vector and $Y \in \mathbb{R}$ be a one-dimensional random variable. Assume a linear-Gaussian model: $Y|\mathbf{x}, \mathbf{w} \sim N(\mathbf{x}^\top\mathbf{w}, 1)$. Suppose that $\mathbf{w} \in \mathbb{R}^d$ is a d-dimensional Gaussian random vector $\mathbf{w} \sim N(0, \mathbf{\Sigma})$, where $\mathbf{\Sigma}$ is a known symmetric positive-definite covariance matrix.

(a) Let us assume that we are given n training data points $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_n, y_n)\}$. Derive the posterior distribution of $\mathbf{w}$ given the training data. What is the MAP estimate of $\mathbf{w}$? Compare this result to the solution you achieve in your homework. Comment on how Tikhonov regularization is a generalization of ridge regression from a probabilistic perspective.

[Hint: You may find the following lemma useful. If the probability density function of a random variable is of the form

$$f(\mathbf{v}) = C \cdot \exp\left\{-\frac{1}{2}\mathbf{v}^\top\mathbf{A}\mathbf{v} + \mathbf{b}^\top\mathbf{v}\right\},$$

where C is some constant to make $f(\mathbf{v})$ integrate to 1 and $\mathbf{A}$ is a symmetric positive definite matrix, then $\mathbf{v}$ is distributed as $N(\mathbf{A}^{-1}\mathbf{b}, \mathbf{A}^{-1})$.

Solution: We'll start the derivation by applying Bayes' Theorem and omitting the constant terms in the denominator which don't depend on $\mathbf{w}$.

$$f(\mathbf{w}|\mathbf{x}_1, \cdots, \mathbf{x}_n, y_1, \cdots, y_n) \propto \left[\prod_{i=1}^n f(y_i|\mathbf{x}_i, \mathbf{w})\right] f(\mathbf{w})$$

Then, we write out the pdf directly.

$$\left[\prod_{i=1}^n f(y_i|\mathbf{x}_i, \mathbf{w})\right] f(\mathbf{w}) \propto \left[\prod_{i=1}^n \exp\left\{-\frac{1}{2}(y_i - \mathbf{x}_i^\top\mathbf{w})^2\right\}\right] \exp\left\{-\frac{\mathbf{w}^\top\mathbf{\Sigma}^{-1}\mathbf{w}}{2}\right\}$$

We from expand out the squared term and combine it with the exponent in the prior, grouping terms with shared factors.

$$\left[\prod_{i=1}^n \exp\left\{-\frac{1}{2}(y_i - \mathbf{x}_i^\top\mathbf{w})^2\right\}\right] \exp\left\{-\frac{\mathbf{w}^\top\mathbf{\Sigma}^{-1}\mathbf{w}}{2}\right\} = \left[\exp\left\{-\frac{1}{2}\sum_{i=1}^n (y_i - \mathbf{x}_i^\top\mathbf{w})^2\right\}\right] \exp\left\{-\frac{\mathbf{w}^\top\mathbf{\Sigma}^{-1}\mathbf{w}}{2}\right\}$$

$$\propto \exp\left\{-\frac{1}{2}[\mathbf{w}^T(\mathbf{X}^\top\mathbf{X}+\boldsymbol{\Sigma}^{-1})\mathbf{w}-2(\mathbf{X}^\top\mathbf{y})^\top\mathbf{w}]\right\},$$

Now, we have $\mathbf{y} \in \mathbb{R}^n$ with ith entry y_i. Based on the lemma we provided above, we can therefore determine the posterior of $\mathbf{w}|\mathbf{x}_1,\cdots,\mathbf{x}_n,y_1,\cdots,y_n$ is $N\Big((\mathbf{X}^\top\mathbf{X}+\boldsymbol{\Sigma}^{-1})^{-1}\mathbf{X}^\top\mathbf{y},(\mathbf{X}^\top\mathbf{X}+\boldsymbol{\Sigma}^{-1})^{-1}\Big)$. We provide a proof of the Lemma below.

Lemma: If the probability density function of a random variable is of the form

$$f(\mathbf{v}) = C\cdot\exp\left\{-\frac{1}{2}\mathbf{v}^\top\mathbf{A}\mathbf{v}+\mathbf{b}^\top\mathbf{v}\right\},$$

where C is some constant to make $f(\mathbf{v})$ integrate to 1, and $\mathbf{A}$ is a symmetric positive definite matrix, then $\mathbf{v}$ is distributed as $N(\mathbf{A}^{-1}\mathbf{b},\mathbf{A}^{-1})$.

The proof of the lemma is as follows:

$$\begin{aligned}
f(\mathbf{v}) &= C\cdot\exp\Big\{-\frac{1}{2}\mathbf{v}^\top\mathbf{A}\mathbf{v}+\mathbf{b}^\top\mathbf{v}\Big\},\\
&= C_1\cdot\exp\Big\{-\frac{1}{2}\Big(\mathbf{v}-(\mathbf{A}^{-1}\mathbf{b})\Big)^\top\mathbf{A}\Big(\mathbf{v}-(\mathbf{A}^{-1}\mathbf{b})\Big)-\frac{1}{2}\mathbf{b}^\top\mathbf{A}^{-1}\mathbf{b}\Big\},\\
&= C_2\cdot\exp\Big\{-\frac{1}{2}\Big(\mathbf{v}-(\mathbf{A}^{-1}\mathbf{b})\Big)^\top\mathbf{A}\Big(\mathbf{v}-(\mathbf{A}^{-1}\mathbf{b})\Big)\Big\},
\end{aligned}$$

which is the density of a Gaussian random variable with mean $\mathbf{A}^{-1}\mathbf{b}$ and covariance matrix $\mathbf{A}^{-1}$.

Now that we have the posterior distribution of $\mathbf{w}$, it is simple to find the MAP estimate. The MAP estimate is simply a maximization over the posterior probability. Since we know the posterior of $\mathbf{w}$ is a Gaussian, we have that the MAP estimate is the mean of that Gaussian. So we have that MAP estimate of $\mathbf{w}$ is $(\mathbf{X}^\top\mathbf{X}+\boldsymbol{\Sigma}^{-1})^{-1}\mathbf{X}^\top\mathbf{y}$.

In ridge regression, we also assume a Gaussian prior on $\mathbf{w}$. However, we constrain the covariance of the prior to be a simple identity scaled by some constant. Tikhonov regularization generalizes this and allows us to select any covariance matrix for our prior on $\mathbf{w}$. Note in particular that this means our prior can involve non-zero covariance terms, or scale variances in different directions differently (e.g., the first entry of $\mathbf{w}$ may have a smaller variance than the second entry; Tikhonov regularization allows us to encode this knowledge in the problem setup. Being able to give priority to certain directions in feature space can be important if we want to be able to take advantage of the regularizing effect of lots of features.

(b) Let us extend this result from the previous part to the case where we introduce observation noise variables Z_i that are not independent across samples, i.e. $\mathbf{Z}$ is not $N(\mathbf{0},\mathbb{I}_n)$ but instead distributed as $N(\boldsymbol{\mu}_z,\boldsymbol{\Sigma}_z)$ for some mean μ_z and some covariance $\boldsymbol{\Sigma}_z$ (still independent of the parameter $\mathbf{w}$). We make the reasonable assumption that the $\boldsymbol{\Sigma}_z$ is invertible. Derive the posterior distribution of $\mathbf{w}$ by appropriately changing coordinates.

(Hint: Write $\mathbf{Z}$ as a function of a standard normal Gaussian vector $\mathbf{V}\sim N(\mathbf{0},\mathbb{I}_n)$ and use the result in (a) for an equivalent model of the form $\widetilde{\mathbf{y}}=\widetilde{\mathbf{X}}\mathbf{w}+\mathbf{V}$.)

Solution: By changing variables, our goal is to reduce to the previous case — with white observation noises.

We want to define a matrix $\boldsymbol{\Sigma}_z^{1/2}$ such that $\boldsymbol{\Sigma}_z^{1/2}(\boldsymbol{\Sigma}_z^{1/2})^\top = \boldsymbol{\Sigma}_z$. According to eigenvalue decomposition $\boldsymbol{\Sigma}_z = \mathbf{U}\Delta\mathbf{U}^\top$. Exploiting the spectral theorem's guarantee of orthonormal eigenvalues and all positive eigenvalues to form the diagonal matrix Δ, we get that we can choose $\boldsymbol{\Sigma}_z^{\frac{1}{2}} = \mathbf{U}\Delta^{\frac{1}{2}}$.

Now, we define $\mathbf{V}$ such that $\mathbf{Z} = \boldsymbol{\mu}_z + \boldsymbol{\Sigma}_z^{\frac{1}{2}}\mathbf{V}$. Then $\mathbf{V} \sim N(\mathbf{0}, \mathbb{I})$ and $\mathbf{V} = \boldsymbol{\Sigma}_z^{-\frac{1}{2}}(\mathbf{Z} - \boldsymbol{\mu}_z)$.

$$
\begin{aligned}
\mathbf{y} &= \mathbf{X}\mathbf{w} + \mathbf{Z} \\
\mathbf{y} - \boldsymbol{\mu}_z &= \mathbf{X}\mathbf{w} + \mathbf{Z} - \boldsymbol{\mu}_z \\
\boldsymbol{\Sigma}_z^{-\frac{1}{2}}(\mathbf{y} - \boldsymbol{\mu}_z) &= \boldsymbol{\Sigma}_z^{-\frac{1}{2}}\mathbf{X}\mathbf{w} + \boldsymbol{\Sigma}_z^{-\frac{1}{2}}(\mathbf{Z} - \boldsymbol{\mu}_z) \\
\boldsymbol{\Sigma}_z^{-\frac{1}{2}}(\mathbf{y} - \boldsymbol{\mu}_z) &= \boldsymbol{\Sigma}_z^{-\frac{1}{2}}\mathbf{X}\mathbf{w} + \mathbf{V}
\end{aligned}
$$

So the problem reduces to the previous part by denoting $\boldsymbol{\Sigma}_z^{-\frac{1}{2}}\mathbf{X}$ as $\widetilde{\mathbf{X}}$ and denoting $\widetilde{\mathbf{y}}$ as $\boldsymbol{\Sigma}_z^{-\frac{1}{2}}(\mathbf{y} - \boldsymbol{\mu}_z)$, which yields the posterior of $\mathbf{w}|\mathbf{x}_1, \cdots, \mathbf{x}_n, y_1, \cdots, y_n$ as a multivariate Gaussian

$$N\left((\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}} + \boldsymbol{\Sigma}^{-1})^{-1}\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{y}}, (\widetilde{\mathbf{X}}^\top\widetilde{\mathbf{X}} + \boldsymbol{\Sigma}^{-1})^{-1}\right)$$

or

$$N\left((\mathbf{X}^\top\boldsymbol{\Sigma}_z^{-1}\mathbf{X} + \boldsymbol{\Sigma}^{-1})^{-1}\mathbf{X}^\top\boldsymbol{\Sigma}_z^{-1}(\mathbf{y} - \boldsymbol{\mu}_z), (\mathbf{X}^\top\boldsymbol{\Sigma}_z^{-1}\mathbf{X} + \boldsymbol{\Sigma}^{-1})^{-1}\right).$$

Notice here how the $\boldsymbol{\Sigma}_z^{-1}$ matrix is now sitting in between the $\mathbf{X}^\top\boldsymbol{\Sigma}_z^{-1}\mathbf{X}$ terms.