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.

DIS2, ©UCB CS 189/289A, Fall 2022. All Rights Reserved. This may not be publicly shared without explicit permission.

(a) What is the MLE of the bias of the coin θ?

(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 $\frac{a}{\theta} - \frac{b}{1-\theta} - \frac{\theta - c}{d}$.

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

(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?

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

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 $N(\mathbf{A}^{-1}\mathbf{b}, \mathbf{A}^{-1})$.

(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* $\widetilde{\mathbf{y}} = \widetilde{\mathbf{X}}\mathbf{w} + \mathbf{V}$*.)*