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 $\hat{\theta}_{\text{MLE}}$?**

(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.

[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 $N(0.8, 1)$ instead?

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

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

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:

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

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 $p(\mathbf{w})$.

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

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

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

What is the constant λ in terms of given quantities?

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.**

(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 $f_{W,Y}(w, y) = f_W(w) f_Y(y)$.

(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?**

(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** $\min_{x:||x||_2=1} ||Ax||_2^2$**?**

(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?**