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

DIS2

1 ℓ_1- and ℓ_2-Regularization

Consider sample points $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$ and associated values $y_1, y_2, \ldots, y_n \in \mathbb{R}$, an $n \times d$ design matrix $X = [X_1 \quad \ldots \quad X_n]^\top$ and an n-vector $y = [y_1 \quad \ldots \quad y_n]^\top$.

For the sake of simplicity, assume (1) that the sample data have been centered (i.e each feature has mean 0) and (2) that the sample data have been whitened, meaning a linear transformation is applied to the original data matrix so that the resulting features have variance 1 and the features are uncorrelated; i.e., $X^\top X = nI$.

For this question, we will not use a fictitious dimension nor a bias term; our linear regression function will output zero for $x = 0$.

Consider linear least-squares regression with regularization in the ℓ_1-norm, also known as Lasso. The Lasso cost function is

$$J(w) = \|Xw - y\|_2^2 + \lambda \|w\|_1$$

where $w \in \mathbb{R}^d$ and $\lambda > 0$ is the regularization parameter. Let $w^* = \operatorname{argmin}_{w \in \mathbb{R}^d} J(w)$ denote the weights that minimize the cost function.

In the following steps, we will explore the sparsity-promoting property of the ℓ_1-norm and compare this with the ℓ_2-norm.

1. We use the notation X_{*i} to denote column i of the design matrix X, which represents the i^{th} feature. Write $J(w)$ in the following form for appropriate functions g and f.

$$J(w) = g(y) + \sum_{i=1}^{d} f(X_{*i}, w_i, y, \lambda)$$

Solution: We expand the objective, and simplify it using the fact that $X^\top X = nI$ since the data is whitened.

$$\begin{aligned} J(w) &= w^\top X^\top X w - 2y^\top X w + \lambda \|w\|_1 + \|y\|^2 \\ &= n\|w\|^2 - 2y^\top X w + \lambda \|w\|_1 + \|y\|^2 \end{aligned}$$

We can write each term in sum form: $n\|w\|^2 = \sum_{i=1}^d n w_i^2$. $\lambda\|w\|_1 = \sum_{i=1}^d \lambda |w_i|$. And $-2y^\top X w = \sum_{i=1}^d -2y^\top X_{*i} w_i$. So the appropriate functions are $g(y) = \|y\|^2$, and

$$f(X_{*i}, w_i, y, \lambda) = n w_i^2 - 2y^\top X_{*i} w_i + \lambda |w_i|$$

2. If $w_i^* > 0$, solve for the optimal value w_i^*. *Hint: use your answer in the previous part.*

   Solution: We want to minimize

   $$-2y^\top X_{*i} w_i + n w_i^2 + \lambda w_i.$$

   Setting the derivative to zero yields

   $$w_i^* = \frac{1}{n}(y^\top X_{*i} - \lambda/2).$$

3. If $w_i^* < 0$, solve for the optimal value w_i^*.

   Solution: We want to minimize

   $$-2y^\top X_{*i} w_i + n w_i^2 - \lambda w_i.$$

   Setting the derivative to zero yields

   $$w_i^* = \frac{1}{n}(y^\top X_{*i} + \lambda/2).$$

4. Considering parts 2 and 3, what is the condition for w_i^* to be zero?

   Solution: w_i^* cannot be positive if $y^\top X_{*i} - \lambda/2 \le 0$, and w_i^* cannot be negative if $y^\top X_{*i} + \lambda/2 \ge 0$. So w_i^* is zero if both are true, i.e., $-\lambda \le 2y^\top X_{*i} \le \lambda$.

5. Now consider ridge regression, which uses the ℓ_2 regularization term $\lambda |w|^2$. How does this change the function $f(\cdot)$ from part 1? What is the new condition in which $w_i^* = 0$? How does it differ from the condition you obtained in part 4?

   Solution: The portion $f(\cdot)$ of the cost function involving w_i is

   $$-2y^\top X_{*i} w_i + n w_i^2 + \lambda w_i^2.$$

   Setting the derivative to zero yields

   $$w_i^* = \frac{y^\top X_{*i}}{n + \lambda}.$$

   Hence w_i^* is zero if $y^\top X_{*i} = 0$. In contrast, $w_i^* = 0$ when $|2y^\top X_{*i}| < \lambda$ in Lasso regression. This is why ℓ_1-norm regularization encourages sparsity.

   Also note that we have shown that whitened training data decouples the features, so that w_i^* is determined by the i^{th} feature alone (i.e., column i of the design matrix X), regardless of the other features. This is true for both Lasso and ridge regression.

2 Probabilistic Interpretation of Lasso

Let's start with the probabilistic interpretation of least squares. Start with labels $y \in \mathbb{R}$, data $\mathbf{x} \in \mathbb{R}^d$, and noise $z \sim \mathcal{N}(0, \sigma^2)$, where $y = \mathbf{w}^\top \mathbf{x} + z$. Recall from lecture that we then have

$$P(y \mid \mathbf{x}, \mathbf{w}, \sigma^2) \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}, \sigma^2)$$

However, maximum likelihood estimates (MLE) can overfit by picking parameters that mirror the training data. To ameliorate this issue, we can assume a Laplace prior on $w_j \sim \text{Laplace}(0, t)$, i.e.

$$P(w_j) = \frac{1}{2t} e^{-|w_j|/t}$$
$$P(\mathbf{w}) = \prod_{j=1}^{D} P(w_j) = \left(\frac{1}{2t}\right)^D \cdot e^{-\sum |w_j|/t}$$

Here, we will see that this modification results in the Lasso objective function.

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

$$\begin{aligned}
\mathbf{w}^* &= \underset{\mathbf{w}}{\text{argmax}}\ L(\mathbf{w}) \\
&= \underset{\mathbf{w}}{\text{argmax}}\ P(Y_1, \ldots, Y_n \mid \mathbf{w}, \mathbf{X}_1, \ldots, \mathbf{X}_n, \sigma^2) \\
&= \underset{\mathbf{w}}{\text{argmax}} \prod_{i=1}^{n} P(Y_i \mid \mathbf{X}_i, \mathbf{w}, \sigma^2).
\end{aligned}$$

When working in a Bayesian framework, we instead focus on the posterior distribution of the parameters conditioned on the data, $P(\mathbf{w} \mid Y_1, \ldots, Y_n, \mathbf{X}_1, \ldots, \mathbf{X}_n, \sigma^2)$. To pick a single model, we can choose the $\mathbf{w}$ that is most likely according to the posterior,

$$\begin{aligned}
\mathbf{w}^* &= \underset{\mathbf{w}}{\text{argmax}}\ P(\mathbf{w} \mid Y_1, \ldots, Y_n, \mathbf{X}_1, \ldots, \mathbf{X}_n, \sigma^2) \\
&= \underset{\mathbf{w}}{\text{argmax}}\ \frac{P(\mathbf{w}, Y_1, \ldots, Y_n \mid \mathbf{X}_1, \ldots, \mathbf{X}_n, \sigma^2)}{P(Y_1, \ldots, Y_n \mid \mathbf{X}_1, \ldots, \mathbf{X}_n, \sigma^2)} \\
&= \underset{\mathbf{w}}{\text{argmax}}\ \frac{P(Y_1, \ldots, Y_n \mid \mathbf{w}, \mathbf{X}_1, \ldots, \mathbf{X}_n, \sigma^2) P(\mathbf{w})}{P(Y_1, \ldots, Y_n \mid \mathbf{X}_1, \ldots, \mathbf{X}_n, \sigma^2)} \\
&= \underset{\mathbf{w}}{\text{argmax}}\ \frac{L(\mathbf{w}) P(\mathbf{w})}{P(Y_1, \ldots, Y_n)} \\
&= \underset{\mathbf{w}}{\text{argmax}}\ L(\mathbf{w}) P(\mathbf{w}) \qquad \text{since } P(Y_1, \ldots, Y_n \mid \mathbf{X}_1, \ldots, \mathbf{X}_n, \sigma^2) \text{ does not depend on } \mathbf{w}.
\end{aligned}$$

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

(a) Write the log-likelihood for this MAP estimate.

Solution:

We start with the likelihood.

$$P(w \mid \mathbf{X}_i, Y_i) \propto \left(\prod_{i=1}^{n} \mathcal{N}(Y_i \mid \mathbf{w}^\top \mathbf{X}_i, \sigma^2) \right) \cdot P(\mathbf{w}) = \left(\prod_{i=1}^{n} \mathcal{N}(Y_i \mid \mathbf{w}^\top \mathbf{X}_i, \sigma^2) \right) \cdot \prod_{j=1}^{D} P(w_j)$$

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

$$\begin{aligned} l(\mathbf{w}) &= \sum_{i=1}^{n} \log \mathcal{N}(Y_i \mid \mathbf{w}^\top \mathbf{X}_i, \sigma^2) + \sum_{j=1}^{D} \log P(w_j) \\ &= \sum_{i=1}^{n} \log \left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(Y_i - \mathbf{w}^\top \mathbf{X}_i)^2}{2\sigma^2} \right) \right) + \sum_{j=1}^{D} \log \left(\frac{1}{2t} \exp\left(\frac{-|w_j|}{t} \right) \right) \\ &= -\sum_{i=1}^{n} \frac{(Y_i - \mathbf{w}^\top \mathbf{X}_i)^2}{2\sigma^2} - \sum_{j=1}^{D} \frac{|w_j|}{t} + n \log \left(\frac{1}{\sqrt{2\pi}\sigma} \right) + D \log \left(\frac{1}{2t} \right) \end{aligned}$$

(b) We already have the log-likelihood for MAP. Show that the MAP you derived in the previous part (in this case, Gaussian noise with a Laplace prior) is equivalent to minimizing the following objective. Additionally, identify the constant λ. Note that $\|\mathbf{w}\|_1 = \sum_{j=1}^{D} |w_j|$.

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

Solution: We first negate the whole expression to change it from a maximization problem to a minimization problem:

$$\begin{aligned} \mathbf{w}^* &= \underset{\mathbf{w}}{\text{argmax}} \; -\sum_{i=1}^{n} \frac{(Y_i - \mathbf{w}^\top \mathbf{X}_i)^2}{2\sigma^2} - \sum_{j=1}^{D} \frac{|w_j|}{t} + n \log \left(\frac{1}{\sqrt{2\pi}\sigma} \right) + D \log \left(\frac{1}{2t} \right) \\ &= \underset{\mathbf{w}}{\text{argmin}} \; \sum_{i=1}^{n} \frac{(Y_i - \mathbf{w}^\top \mathbf{X}_i)^2}{2\sigma^2} + \sum_{j=1}^{D} \frac{|w_j|}{t} - n \log \left(\frac{1}{\sqrt{2\pi}\sigma} \right) - D \log \left(\frac{1}{2t} \right) \end{aligned}$$

Now, we drop constants from the expression above and multiply everything else with $2\sigma^2$ — these transformations will not affect the argmin of the objective.

$$\begin{aligned} \mathbf{w}^* &= \underset{\mathbf{w}}{\text{argmin}} \; \sum_{i=1}^{n} (Y_i - \mathbf{w}^\top \mathbf{X}_i)^2 + \frac{2\sigma^2}{t} \sum_{j=1}^{D} |w_j| \\ &= \underset{\mathbf{w}}{\text{argmin}} \; \sum_{i=1}^{n} (Y_i - \mathbf{w}^\top \mathbf{X}_i)^2 + \lambda \|\mathbf{w}\|_1 \\ &= \underset{\mathbf{w}}{\text{argmin}} \; J(\mathbf{w}) \end{aligned}$$

for $\lambda = \frac{2\sigma^2}{t}$.

3 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:

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

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 (probability of flipping heads in a single trial). 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.

(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.

[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.