6.7900 Fall 2024: Lecture Notes 5

1 Recap

Proposition: Consider regression with $X = \mathbb{R}^D, Y = \mathbb{R}$ and square loss $L(a, g) = (a - g)^2$, then decision rule minimizes risk of a new point $h(x) = \mathbb{E}[Y|X = x]$.

We are not aware of future data points, but by assuming past data and future data are iid, we can take use of training data. A maximum likelihood approach would model

$$p(y|x, \theta, \sigma^2) = \mathcal{N}(y|\theta^T x, \sigma^2)$$

We can also just minimize empirical risk over all decision rules, but this does not generalize well. Thus, we limit the decision rules to only linear predictors $h(x) = \theta^T x$.

In both of the above cases, we get

$$\hat{\theta} = \arg\min_{\theta}(X\theta - Y)^T((X\theta - Y))$$

and we derived the close form solution when $N > D$ and X is full rank.

2 Problems of Linear Regression

Visualizing the full-rank case See slides for figures of visualization. We see that in the given example, the RSS objective is strictly convex in the 1 dimensional case and converges to a single global minimum. However, it could be that the RSS (residual sum of squares, see lecture 4) objective is not strictly convex and there are multiple minima.

What could go wrong? Consider the case where two or more features are perfectly collinear. There isn't a unique best hyperplane; in fact, there are infinitely many optimal solutions. In this situation, X is no longer full rank.

Exercise: What happens if y has noise in it? Does that change the situation?

Assume that we have a tiny and meaningless noise in x. Now there exists a unique hyperplane, but the unique solution is meaningless. See slides for a visualization of this scenario. An example could be predicting with both credit card limit and credit rating as two features, which may be perfectly or imperfectly collinear. Recall the Gaussian example with "flat" likelihood from lecture 2.

Exercise: Will a more complex model solve this dilemma? What about more data? What is a correct solution?
Hint: Consider not only the mathematical side of the problem, but also the practical application of it.

3 Bayes & Multivariate Gaussians

In this section, we will go back to the case where data is just $\{y^{(n)}\}_{n=1}^{N}$ and develop a Bayesian inference for multivariate Gaussians, then apply this idea to linear regression.

We take a D_y dimensional label $y^{(n)} = [y_1^{(n)}, ..., y_{D_y}^{(n)}]^T$, and suppose that we posit a Gaussian likelihood $y^{(n)} \sim \mathcal{N}(\mu, \Sigma)$ with $\mu \in \mathbb{R}^{D_y}$ and Σ positive definite.

Observation: We have a special case when $\Sigma = \sigma^2 I_{D_y \times D_y}$. This is called a "normal means model". We can get strictly lower risk by estimating parameters jointly in a Bayes-inspired procedure rather than by using separate MLEs (Stein's phenomenon).

We treat Σ as fixed to avoid overload of notations:

$$p(y^{(n)}|\mu) = \frac{1}{\sqrt{(2\pi)^{D_y}|\Sigma|}} e^{-0.5(y^{(n)}-\mu)^T\Sigma^{-1}(y^{(n)}-\mu)}$$

The likelihood can be expressed as

$$p(y^{(n)}|\mu) \propto_\mu e^{-0.5(y^{(n)}-\mu)^T\Sigma^{-1}(y^{(n)}-\mu)}$$

We have the conjugate prior

$$p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$$

In practice, the hyperparameters can be chosen based on domain information. For example, if I want to know the PM2.5 value at a sensor, I can take the NYC's reported value of 0 to 117 μgm^{-3}.

The posterior for one data point can be written as

$$p(\mu|y^{(1)}) \propto_\mu p(y^{(1)}|\mu)p(\mu|\mu_0, \Sigma_0)$$

$$p(\mu|y^{(1)}) \propto_\mu e^{-0.5(y^{(1)}-\mu)^T\Sigma^{-1}(y^{(1)})-\mu}$$

$$p(\mu|y^{(1)}) = \mathcal{N}(\mu|\mu_1, \Sigma_1) \propto_\mu e^{-0.5(\mu-\mu_1)^T\Sigma_1^{-1}(\mu-\mu_1)}$$

We would like to solve for the mean and variance of the posterior. So for any μ, we have

$$\mu^T\Sigma_1^{-1}\mu = \mu^T\Sigma^{-1}\mu + \mu^T\Sigma_0^{-1}\mu$$

$$\Sigma_1^{-1} = \Sigma^{-1} + \Sigma_0^{-1}$$

$$\Sigma_1 = (\Sigma^{-1} + \Sigma_0^{-1})^{-1}$$

We are allowed to take the inverse here because sum of positive definite matrices is also positive definite. For future reference, note that the inverse of a covariance matrix is called the precision matrix. We also have

$$\mu_1^T\Sigma_1^{-1}\mu = (y^{(1)})^T\Sigma^{-1}\mu + \mu_0^T\Sigma_0^{-1}\mu$$

$$\Sigma_1^{-1}\mu_1 = \Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0$$

$$\mu_1 = \Sigma_1(\Sigma^{-1}y^{(1)} + \Sigma_0^{-1}\mu_0)$$

Multiple data points We have:

- Likelihood: $p(y^{(m)}|\mu) = \mathcal{N}(y^{(n)}|\mu, \Sigma)$
- Conjugate prior: $p(\mu) = \mathcal{N}(\mu|\mu_0, \Sigma_0)$
- Posterior for one datapoint: Σ_1, μ_1

We start by solving the posterior for two datapoints $p(\mu|\{y^{(n)}\}_{n=1}^2) = \mathcal{N}(\mu|\mu_2, \Sigma_2)$

$$\Sigma_2^{-1} = \Sigma^{-1} + \Sigma_1^{-1} = 2\Sigma^{-1} + \Sigma_0^{-1}$$

$$\Sigma_2^{-1}\mu_2 = \Sigma^{-1}y^{(2)} + \Sigma_1^{-1}\mu_1 = \Sigma^{-1}\sum_{n=1}^{2} y^{(n)} + \Sigma_0^{-1}\mu_0$$

Recursively, we can obtain the posterior for N data points $p(\mu|\{y^{(n)}\}_{n=1}^N) = \mathcal{N}(\mu|\mu_N, \Sigma_N)$

$$\Sigma_N^{-1} = N\Sigma^{-1} + \Sigma_0^{-1}$$

$$\Sigma_N^{-1}\mu_N = \Sigma^{-1}\sum_{n=1}^{N} y^{(n)} + \Sigma_0^{-1}\mu_0$$

Let's perform a sanity check on the above results under the one label case where $D_y = 1$. We get

$$\sigma_N^2 = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

$$\mu_N = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}(\frac{1}{\sigma^2}\sum_{n=1}^{N} y^{(n)} + \frac{1}{\sigma^2}\mu_0)$$

Exercise: Check what happens when $N = 0$. What happens if $N \to \infty$?

4 Bayesian Linear Regression

Let's look back to linear regression with Bayesian likelihood $\mathcal{N}(y^{(n)}|\theta^T x^{(n)}, \sigma^2)$ with the assumption that σ is known. We implicitly condition on x for all expressions below:

$$p(y^{(1)}|\theta) \propto_\theta e^{-\frac{1}{2\sigma^2}(y^{(1)} - \theta^T x^{(1)})^2}$$

Our conjugate prior is $p(\theta) = \mathcal{N}(\theta|\mu_0, \Sigma_0)$, and our posterior $p(\theta|y^{(1)}) = \mathcal{N}(\theta|\mu_1, \Sigma_1)$. Following a similar procedure from above, we get

$$\Sigma_1^{-1} = \Sigma_0^{-1} + (\sigma^2)^{-1} x^{(-1)} (x^{(1)})^T$$

$$\Sigma_1^{-1}\mu_1 = \Sigma_0^{-1}\mu_0 + (\sigma^2)^{-1} x^{(1)} y^{(1)}$$

Exercise: Check the above result for Σ_1^{-1}.

Recursively, we can get the result for N data points:

$$\Sigma_N^{-1} = \Sigma_0^{-1} + (\sigma^2)^{-1} \sum_{n=1}^{N} x^{(n)} (x^{(n)})^T = \Sigma_0^{-1} + (\sigma^2)^{-1} X^T X$$

$$\Sigma_N^{-1}\mu_N = \Sigma_0^{-1}\mu_0 + (\sigma^2)^{-1} X^T Y$$