LECTURE 1/30

Scribe Notes

Contents

1 Review From Previous Lecture

Last time, we discussed two views on regression: optimization and projection.

1.1 Optimization

Let's examine the first viewpoint: *optimization*. First, we choose a loss function to optimize.

$$\mathcal{L}_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2$$

We choose this specific loss function because it is easy to solve (with the derivation, the 1/2 constant goes away). But the main takeaway is that there exists a choice in what specific loss function gets used. There is tension between choosing a loss function that the real world needs vs. what is easy to implement.

Story of the Day: (Mating system and brain size in bats)

In the real world, understanding the sources of errors in a model is imperative. Consider a study conducted in 2006, where researchers sought to examine the relationship between the size of bats' testes and their brain size. Initially, the data suggested a positive correlation, indicating that bats with larger testes also had larger brains. However, this correlation was confounded by the variable of bat size, as larger bats naturally tend to have larger body parts, including brains and testes.

After correcting for body mass, the correlation flips. But there were still residual errors in the regression model. Upon closer analysis, the researchers discovered that the mating system of bats played a crucial role. Bats engaged in monogamous relationships exhibited larger brains, revealing an additional factor influencing brain size beyond body mass alone. This study underscores the importance of considering confounding variables, such as body size, and additional factors like mating systems when constructing a model.

1.2 Projection

The second view that we have talked about is *projection*. The following is our matrix equation for our predictions $\hat{\mathbf{y}}$.

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}$$

As a sanity check, we can consider the dimensions of each of the matrices.

$$\begin{gathered}\hat{\mathbf{y}} \text{ is } N \text{ dimensional} \\ \mathbf{X} \text{ is } N \text{ x } D \text{ dimensional} \\ \mathbf{w} \text{ is } D \text{ dimensional}\end{gathered}$$

Thus, we can view the $\mathbf{X}$ matrix as D columns.

For intuition with the projection view, let us consider the case where we set $D = 2$. Then, we have two vector columns, $\mathbf{x}_{d=1}$ and $\mathbf{x}_{d=2}$.

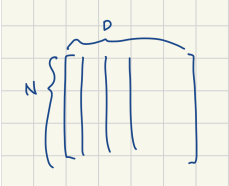


Figure 1: Our X matrix.

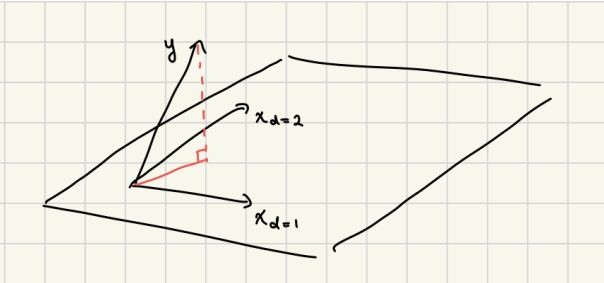


Figure 2: Example projection.

The two vectors $\mathbf{x}_{d=1}$ and $\mathbf{x}_{d=2}$ can only create lines on the plane, so we project $\mathbf{y}$ onto the plane to represent our prediction $\hat{\mathbf{y}}$. By linear algebra properties, we know the projection of $\mathbf{y}$ onto $\mathbf{x}_d$ is

$$\hat{\mathbf{y}}_d = \frac{\mathbf{x}_d}{\langle \mathbf{x}_d, \mathbf{x}_d \rangle} \langle \mathbf{x}_d, y \rangle$$

Recall from last lecture the following formula for our y prediction matrix:

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

The terms from the projection formula match up with the terms in our prediction matrix.

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

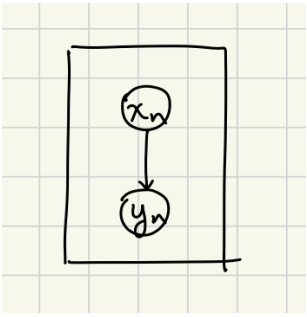
Therefore, intuitively we can think about our prediction matrix $\hat{\mathbf{y}}$ as the projection of $\mathbf{y}$ onto $\mathbf{x}$.

2 Probabilistic View

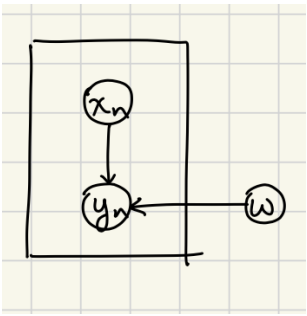
In today's lecture, we will introduce a third view of regression: *probabilistic*. With this probabilistic view, we will tell the story of how we are generating our y's and by telling this story, we will expose the assumptions in our model.

2.1 Probabilistic Graphical Models

We have our $\mathbf{x}_n$'s that are used to generate our $\mathbf{y}_n$'s. Since there are n datapoints, we put the $\mathbf{x}_n$'s and $\mathbf{y}_n$'s inside the box, indicating that there are multiple $\mathbf{x}_n$'s and $\mathbf{y}_n$'s.



We have our weights that help generate $\mathbf{y}_n$'s. Since the weights are the same for every pair of $(\mathbf{x}_n, \mathbf{y}_n)$, so we draw the $\mathbf{w}$ outside of the box.



In our probabilistic view, we add noise in our generation of $\mathbf{y}_n$'s. The formula for $\mathbf{y}_n$ is the following

$$\mathbf{y}_n = \mathbf{w}\mathbf{x}_n + \epsilon_n \text{ where } \epsilon_n \sim N(0, \sigma^2).$$

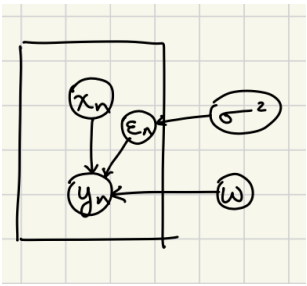
Similar to our weights, the variance of our Gaussian distribution is a hyperparameter that does not change, so σ^2 goes outside of the box. But we have an ϵ for each generation of $\mathbf{y}_n$, so the ϵ belongs inside the box.

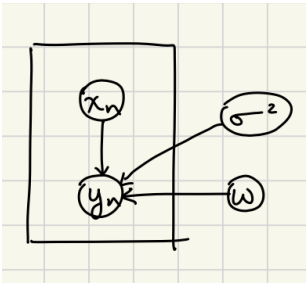
We will often seen this diagram represented without the ϵ_n, and instead a direct arrow from σ^2 to $\mathbf{y}_n$.

2.2 Maximizing Likelihood

The likelihood of a model is the probability of our given data occurring given a model.

$$\Pr(\text{data} \mid \text{model}) = \prod_{n=1}^{N} P(\mathbf{y}_n \mid \mathbf{x}_n, \mathbf{w}, \sigma^2)$$





We take the log on both sides, giving us the benefit of working with summation instead of a product.

$$\log \Pr(\text{data} \mid \text{model}) = \textstyle\sum_{n=1}^{N} \log P(\mathbf{y}_n \mid \mathbf{x}_n, \mathbf{w}, \sigma^2)$$

Notice that because $\mathbf{y}_n = \mathbf{w}\mathbf{x}_n + \epsilon_n$ and $\epsilon_n \sim N(0, \sigma^2)$, then $\mathbf{y}_n \sim N(\mathbf{x}_n\mathbf{w}, \sigma^2)$. Thus, we can use the Gaussian distribution PDF to get

$$\log \Pr(\text{data} \mid \text{model}) = \textstyle\sum_{n=1}^{N} \log[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(\mathbf{y}_n - \mathbf{x}_n\mathbf{w})^2\}].$$

By log and sum properties, we can simplify the right-hand expression.

$$\log \Pr(\text{data} \mid \text{model}) = \textstyle\sum_{n=1}^{N} \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(\mathbf{y}_n - \mathbf{x}_n\mathbf{w})^2$$
$$\log \Pr(\text{data} \mid \text{model}) = N \log \frac{1}{\sqrt{2\pi\sigma^2}} + \frac{1}{2\sigma^2} \textstyle\sum_{n=1}^{N}(\mathbf{y}_n - \mathbf{x}_n\mathbf{w})^2$$

Now our goal is to maximize the log likelihood, which in turn maximizes the likelihood. In order to maximize with respect to σ^2, we find the derivative with respect to σ^2, set the expression to zero, and solve for σ^2.

Taking the derivative with respect to σ^2 gives us

$$0 = N(\tfrac{-1}{2})(\tfrac{1}{\sigma^2}) + \tfrac{1}{2}(\tfrac{1}{\sigma^2})^2 \textstyle\sum_{n=1}^{N}(\mathbf{y}_n - \mathbf{x}_n\mathbf{w})^2.$$

Then, solving for σ^2, we find

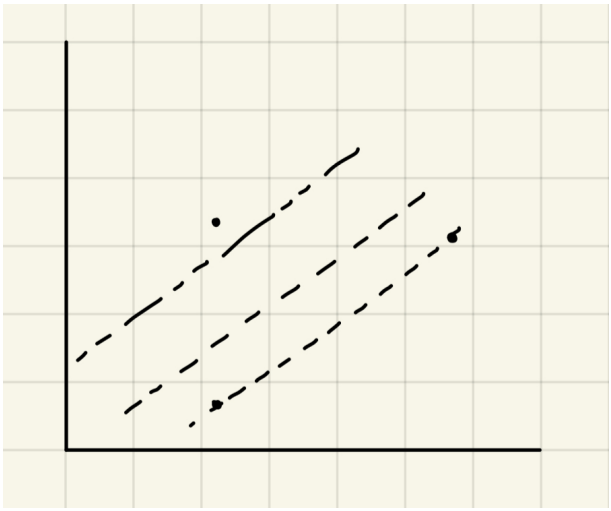
$$\boxed{\sigma^2_{\text{ML}} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{y}_n - \mathbf{x}_n\mathbf{w})^2.}$$

2.3 Concept Check

What if we used a different distribution for noise?

Laplace($\lambda = 1$)	**Gaussian($\sigma^2 = 1$)**	**Student T**
$\Pr(\epsilon) \propto \exp(-\lambda\lvert\epsilon\rvert)$	$\Pr(\epsilon) \propto \exp(-\frac{1}{2}\epsilon^2)$	$\Pr(\epsilon) \propto \exp(1+\epsilon^2)^{-1}$

Which of each distribution would correspond with each dotted line model of the data-points?



Solution:

The regression produced with a Laplace distribution for noise encourages residuals to approach zero. To illustrate, consider an exponential function, where the disparity in x^2 between x values of 10000 and 9999 is substantial, while the difference for x^2 values of 2 to 1 is comparatively minor. In contrast, an absolute value function maintains a consistent difference between two outputs as the distance to zero decreases, consistently rewarding proximity to zero.

Consequently, the Laplace distribution rewards a regression line that attains zero residuals for the provided points. Thus, the line that has zero residuals with the two points will be rewarded with the Laplace distribution. The next closest line to the two points would be generated by a regression with Student T distribution noise. The line closest to the outlier would be from the Gaussian noise model.

The main takeaway here is that different distributions for noise will effect the model. The Gaussian distribution is commonly used in linear regression. The likelihood function assumes that the errors are normally distributed.

The use of the Laplace distribution for noise results in a model that is more robust to outliers. The predicted values may exhibit sparsity, and the model may be less sensitive to extreme observations.

Similar to the Laplace distribution, the use of Student's t-distribution results in a more robust model. It allows for the estimation of degrees of freedom, which affects the tails of the distribution.