

DSC 140A

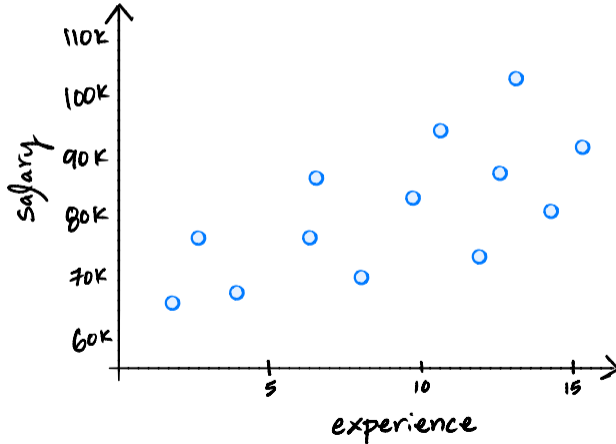
Probabilistic Modeling & Machine Learning

Lecture 16 | Part 1

Recall: Regression

Recall

- ▶ We have seen the problem of regression.



Recall

- ▶ Introduced **empirical risk minimization (ERM)**:
- ▶ Step 1: choose a **hypothesis class**
 - ▶ Let's assume we've chosen linear predictors
- ▶ Step 2: choose a **loss function**
 - ▶ Used square loss
- ▶ Step 3: minimize **expected loss (empirical risk)**
 - ▶ MSE (Mean Squared Error)

Recall: Least Squares

- ▶ Goal: fit a function of the form $H(\vec{x}; \vec{w}) = \text{Aug}(\vec{x}) \cdot \vec{w}$
- ▶ In (ordinary) least squares regression, we **minimized** the **mean squared error**:

$$\vec{w}^* = \arg \min_{\vec{w}} \frac{1}{n} \sum_{i=1}^n (H(\vec{x}^{(i)}; \vec{w}) - y_i)^2$$

- ▶ **Solution:** $\vec{w}^* = (X^T X)^{-1} X^T \vec{y}$

Observation

- ▶ This the “curve fitting” approach to regression.
- ▶ I.e., find a “line of best fit”.
- ▶ There was no consideration of the (random) process that generated the data.

Today

- ▶ Take a probabilistic approach to regression.

DSC 140A

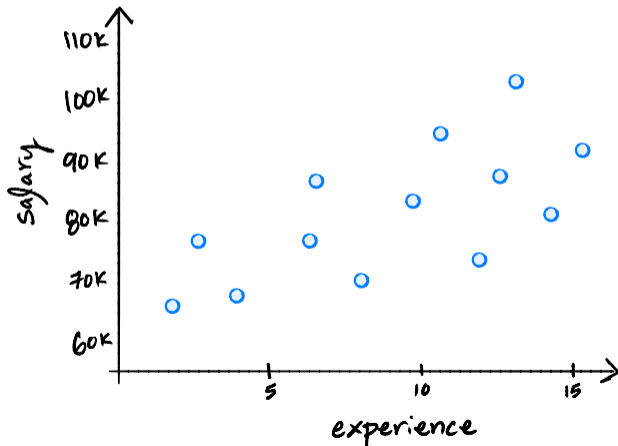
Probabilistic Modeling & Machine Learning

Lecture 16 | Part 2

Probabilistic View of Regression

Probabilistic View of Regression

- **Note:** There is **uncertainty** in the salary.

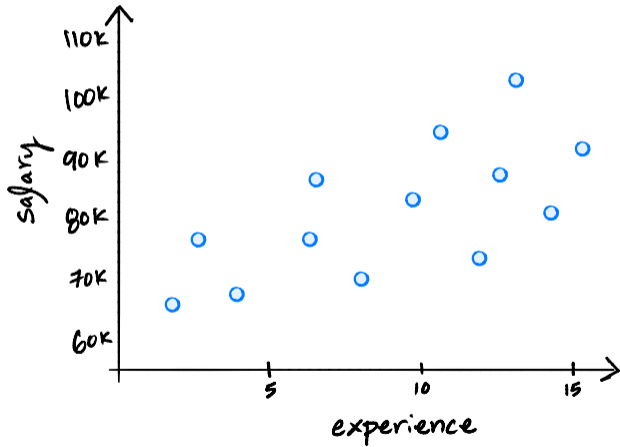


Modeling Uncertainty

- ▶ We can model this uncertainty using probability.

$$\text{Salary} = w_0 + w_1 \times (\text{Experience}) + \varepsilon$$

- ▶ Here, ε is the (random) **error**.
- ▶ What is a reasonable choice of **distribution** for ε ?



Error Distribution

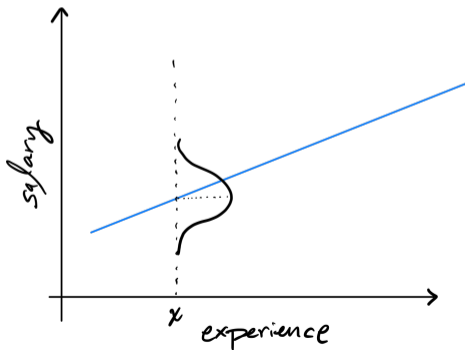
- ▶ It is reasonable to assume that the error distribution is:
 - ▶ **Symmetric:** equally likely to predict high or low
 - ▶ **Centered at zero:** and decreasing as we move away
- ▶ The **Gaussian distribution** (with mean 0) satisfies this.¹

¹In fact, if we add one more constraint, it is the *only* density that satisfies this. See: *deriving the Gaussian error function*.

Modeling Uncertainty

- ▶ Assuming a Gaussian (Normal) distribution:

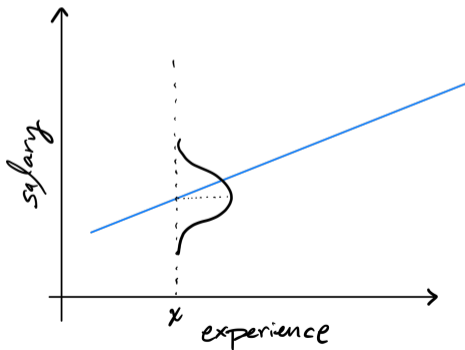
$$\text{Salary} = w_0 + w_1 \times (\text{Experience}) + \underbrace{\mathcal{N}(0, \sigma^2)}_{\varepsilon}$$



Modeling Uncertainty

- ▶ Equivalently:

$$\text{Salary} \sim \mathcal{N}(w_0 + w_1 \times \text{Experience}, \sigma^2)$$



In General

- ▶ In general:

$$Y \sim \mathcal{N}(\text{Aug}(\vec{x}) \cdot \vec{w}, \sigma^2)$$

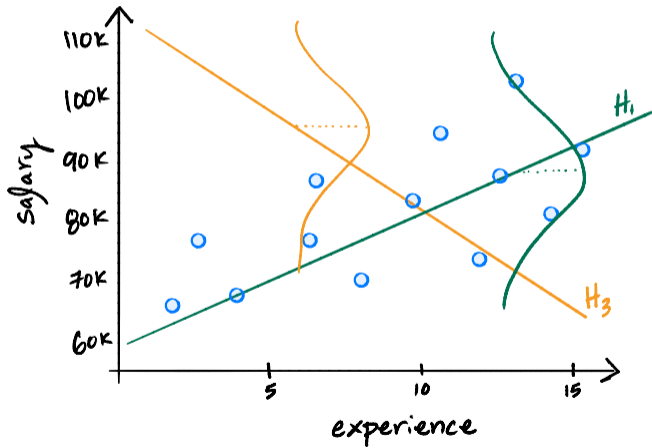
- ▶ That is: for any feature vector \vec{x} , the target Y is drawn from a Gaussian centered at $\text{Aug}(\vec{x}) \cdot \vec{w}$.

Estimating Parameters

- ▶ We assume the model:

$$\text{Salary} \sim \mathcal{N}(w_0 + w_1 \times \text{Experience}, \sigma^2)$$

- ▶ Given some data, what parameters generated it?
 - ▶ What were w_0 , w_1 , σ ?
- ▶ **Estimate** them with maximum likelihood?



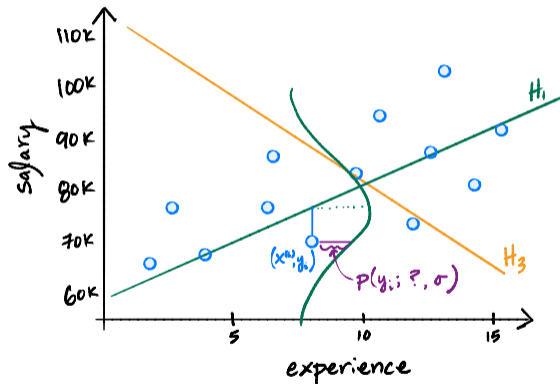
Likelihood

- ▶ Let $p(y; \mu, \sigma)$ be the Gaussian pdf:

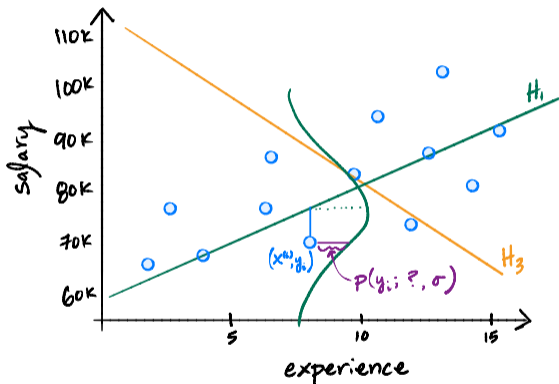
$$p(y; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\mu)^2/(2\sigma^2)}$$

- ▶ We observe a data set $\{(\vec{x}^{(i)}, y_i)\}$.
- ▶ What is the likelihood of a choice of parameters \vec{w}, σ , with respect to the data?

Likelihood wrt a Point



Likelihood wrt a Point



- ▶ $p(y_i; w_0 + w_1 x^{(i)}, \sigma)$ measures likelihood with respect to $(x^{(i)}, y_i)$.

Likelihood

- ▶ In general,

$$p(y_i; \text{Aug}(\vec{x}^{(i)}) \cdot \vec{w}, \sigma)$$

measures likelihood with respect to single data point $(\vec{x}^{(i)}, y_i)$.

- ▶ Likelihood with respect to data set:

$$L(\vec{w}, \sigma) = \prod_{i=1}^n p(y_i; \text{Aug}(\vec{x}^{(i)}) \cdot \vec{w}, \sigma)$$

Log-Likelihood

Compute the log-likelihood from
 $\prod_{i=1}^n p(y_i; \text{Aug}(\vec{x}^{(i)}) \cdot \vec{w}, \sigma).$

Log-Likelihood

- ▶ The log-likelihood is:

$$\tilde{L}(\vec{w}, \sigma) = -\frac{1}{2\sigma^2} \sum_{i=1}^n (\text{Aug}(\vec{x}^{(i)}) \cdot \vec{w} - y_i)^2 + \frac{n}{2} \ln \frac{1}{\sigma^2} - \frac{n}{2} \ln(2\pi)$$

- ▶ We want to **maximize** this quantity.

Claim 1

$$\begin{aligned} & \arg \max_{\vec{w}} \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (\text{Aug}(\vec{x}^{(i)}) \cdot \vec{w} - y_i)^2 + \frac{n}{2} \ln \frac{1}{\sigma^2} - \frac{n}{2} \ln(2\pi) \right] \\ & = \\ & \arg \max_{\vec{w}} \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (\text{Aug}(\vec{x}^{(i)}) \cdot \vec{w} - y_i)^2 \right] \end{aligned}$$

Claim 2

$$\begin{aligned} & \arg \max_{\vec{w}} \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (\text{Aug}(\vec{x}^{(i)}) \cdot \vec{w} - y_i)^2 \right] \\ & = \\ & \arg \max_{\vec{w}} \left[-\frac{1}{n} \sum_{i=1}^n (\text{Aug}(\vec{x}^{(i)}) \cdot \vec{w} - y_i)^2 \right] \end{aligned}$$

Claim 3

$$\begin{aligned} & \arg \max_{\vec{w}} \left[-\frac{1}{n} \sum_{i=1}^n (\text{Aug}(\vec{x}^{(i)}) \cdot \vec{w} - y_i)^2 \right] \\ & = \\ & \arg \min_{\vec{w}} \left[\frac{1}{n} \sum_{i=1}^n (\text{Aug}(\vec{x}^{(i)}) \cdot \vec{w} - y_i)^2 \right] \end{aligned}$$

- ▶ That is, minimize the **mean squared error**.

Main Idea

Maximizing the likelihood of \vec{w} with respect to the data (assuming Gaussian error term) is **equivalent** to minimizing mean squared error.

Solution

- ▶ The maximum likelihood estimate for \vec{w} is therefore:

$$\vec{w}^* = (X^T X)^{-1} X^T \vec{y}$$

- ▶ That is, the exact same as we obtained by empirical risk minimization with the square loss.

DSC 140A

Probabilistic Modeling & Machine Learning

Lecture 16 | Part 3

A Probabilistic View of Regularization

Probabilistic View

- ▶ We've assumed Nature generates data from some underlying distribution that **we don't know**.
- ▶ We assume it has some shape.
 - ▶ E.g., Gaussian.
- ▶ Given some data, we use it to **estimate** the parameters of this distribution.

Two Approaches

- ▶ We've seen one approach: **maximum likelihood**.
 - ▶ Find parameters \vec{w} that maximize $p(\vec{x} | \vec{w})$.
 - ▶ As we'll see, this tends to **overfit**.
- ▶ There's another approach: **Bayesian**.

Example

- ▶ You find a coin on the street.
- ▶ **Assumption:** when you flip it, Nature generates “Heads” with some unknown probability θ .
- ▶ You flip a coin 10 times, getting 7 heads.
- ▶ What was θ ?

Example: the MLE Approach

- ▶ The MLE for θ is:

$$\theta_{\text{MLE}} = \frac{\# \text{ heads}}{\# \text{ flips}} = \frac{7}{10} = 70\%$$

- ▶ Seems high! A coin on the street is probably fair.
- ▶ This is **overfitting**.

Example: the Bayesian Approach

- ▶ In the MLE approach, we try to maximize $p(\text{data} | \theta)$.
- ▶ In the **Bayesian** approach, we treat θ as random.
- ▶ Try to maximize $p(\theta | \text{data})$.

Example: the Bayesian Approach

- ▶ Using Bayes' Rule:

$$p(\theta | \text{data}) \propto p(\text{data} | \theta) \underbrace{p(\theta)}_{\text{prior}}$$

- ▶ We choose a **prior** $p(\theta)$.
 - ▶ This encodes our **prior belief** about θ .
 - ▶ E.g., we might make $p(\theta)$ a Gaussian centered at $1/2$.

Example: the Bayesian Approach

- ▶ Bayesian estimate for θ is the one maximizing:

$$\underbrace{p(\text{data} \mid \theta)}_{\text{likelihood}} \times \underbrace{p(\theta)}_{\text{prior}}$$

- ▶ When we have little data, the maximum of $p(\theta \mid \text{data})$ is close to the maximum of $p(\theta)$.
 - ▶ With 7 out of 10 heads, estimate is 0.52.
- ▶ With more data, maximum is closer to MLE:
 - ▶ With 700 out of 1000 heads, estimate is 0.68.

Regression

- ▶ We derived least squares regression using MLE; what is the Bayesian view?
- ▶ Imagine we have yet to see the data.
- ▶ There is no reason to believe that a given weight w_i is positive or negative.
- ▶ We believe it is more likely to be small (close to zero) than large.

A Prior on Weights

- ▶ This **prior belief** is captured by assuming:

$$w_i \sim \mathcal{N}(0, s^2)$$

- ▶ We are adopting a **Bayesian** view of probability; it expresses level of belief.

A Prior on Weights

- ▶ If each weight has distribution $\mathcal{N}(0, s^2)$, then:

$$\vec{w} \sim \mathcal{N}(\vec{0}, s^2 \cdot I)$$

- ▶ That is, the distribution of \vec{w} has density:

$$p_{\vec{w}}(\vec{w}) = \frac{1}{(2\pi s^2)^{d/2}} e^{-\frac{1}{2} \frac{\|\vec{w}-\vec{0}\|^2}{s^2}}$$

Distribution of \vec{w}

- ▶ Using Bayes' Rule:

$$p_{\vec{w}}(\vec{w} | \vec{x}, y) \propto p_y(y | \vec{w}, \vec{x})p_{\vec{w}}(\vec{w})$$

- ▶ What is the most probable value of \vec{w} ?

$$\begin{aligned}\arg \max_{\vec{w}} [p_{\vec{w}}(\vec{w} | \vec{x}, y)] &= \arg \max_{\vec{w}} [p_y(y | \vec{w}, \vec{x}) p_{\vec{w}}(\vec{w})] \\ &= \arg \max_{\vec{w}} \ln [p_y(y | \vec{w}, \vec{x}) p_{\vec{w}}(\vec{w})] \\ &= \arg \max_{\vec{w}} [\ln p_y(y | \vec{w}, \vec{x}) + \ln p_{\vec{w}}(\vec{w})] \\ &= \arg \min_{\vec{w}} [-\ln p_y(y | \vec{w}, \vec{x}) - \ln p_{\vec{w}}(\vec{w})] \\ &= \arg \min_{\vec{w}} [\text{MSE}(\vec{w}) - \ln p_{\vec{w}}(\vec{w})]\end{aligned}$$

Deriving the Regularizer

► Since

$$p_{\vec{w}}(\vec{w}) = \frac{1}{(2\pi s^2)^{d/2}} e^{-\frac{1}{2} \frac{\|\vec{w}-\vec{0}\|^2}{s^2}}$$

we have:

$$-\ln p_{\vec{w}}(\vec{w}) = c + \frac{1}{2s^2} \|\vec{w}\|^2$$

► So

$$\arg \min_{\vec{w}} [\text{MSE}(\vec{w}) - \ln p_{\vec{w}}(\vec{w})] = \arg \min_{\vec{w}} \left[\text{MSE}(\vec{w}) + \underbrace{\frac{1}{2s^2}}_{\lambda} \|\vec{w}\|^2 \right]$$

Recall: Ridge Regression

- ▶ In **ridge regression**, we added a regularization term: $\|\vec{w}\|^2$.

$$\vec{w}^* = \arg \min_{\vec{w}} \frac{1}{n} \sum_{i=1}^n (H(\vec{x}^{(i)}; \vec{w}) - y_i)^2 + \lambda \|\vec{w}\|^2$$

- ▶ **Solution:** $\vec{w}^* = (X^T X + n\lambda I)^{-1} X^T \vec{y}$
- ▶ Helps control overfitting.

Main Idea

Minimizing the $\|\vec{w}\|^2$ -regularized mean squared error (**ridge regression**) is **equivalent** to placing a $\mathcal{N}(0, s^2)$ prior on each weight and maximizing $p_{\vec{w}}(\vec{w} | \vec{x}, y)$.