

CS189: Intro to Machine Learning

Summer 2018

Lecture 4: Maximum a priori for regression

Josh Tobin
UC Berkeley EECS

Announcements

- HW1 due tonight at 10pm on gradescope.
- Extra office hours 4-6pm in Cory 540AB

OpenAI bot beats humans in Dota



- Challenging problem
 - High-dimensional continuous observation and action spaces
 - Long time horizons (strategy!)
- Simple solution
 - Scale up existing RL method (PPO = policy gradient variant) to 256 GPUs and 128,000 CPU cores

<https://blog.openai.com/openai-five/>

Outline for today

- Review of maximum likelihood estimation
- Maximum a priori
- Bias-variance tradeoff

Outline for today

- **Review of maximum likelihood estimation**
- Maximum a priori
- Bias-variance tradeoff

Four levels for ML problems

1. Data & application
2. Model
3. Optimization problem
4. Optimization algorithm

X, y . Goal: predict y .

Linearly parameterized

$$\min_w ||Xw - y||_2^2$$

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

Justifying least squares

1. Define a probabilistic model for our data
2. Define what it means for a model to be “good” probabilistically (i.e., MLE)
3. Do some math to show that the “best” linear model probabilistically corresponds to the OLS estimate

A probabilistic model for supervised learning

$$y_i = f(\vec{x}_i) + z_i$$

What is z?

- Mean = 0
- All z_i are *independent*
- Each z_i comes from the same distribution (*identically distributed*)

i.i.d
assumptions
(independent
identically
distributed)

A probabilistic model for supervised learning

$$y_i = f(\vec{x}_i) + z_i$$

What is z?

- zero-mean, i.i.d
- Assume a Gaussian distribution

$$z_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

Therefore:

$$y_i \stackrel{\text{iid}}{\sim} \mathcal{N}(f(\vec{x}_i), \sigma^2)$$

Maximum Likelihood Estimation

f_θ = Family of models, e.g., $f_\theta(x) = \theta^T x$

Each theta gives us an implied value for the probability
(i.e., *likelihood*) of the data we observed

$$\mathcal{L}(\theta; \mathcal{D}) = p(\text{data} = \mathcal{D} \mid \text{model} = f_\theta)$$

The “best” model is the one that maximizes the likelihood

$$\hat{\theta}_{\text{MLE}} = \operatorname{argmax}_\theta \mathcal{L}(\theta; \mathcal{D})$$

MLE justifies OLS

$$\hat{\theta}_{\text{MLE}} = \operatorname{argmin}_{\theta} \sum_{i=1}^n (y_i - f_{\theta}(x_i))^2$$

Linear regression model:

$$\hat{\theta}_{\text{MLE}} = \operatorname{argmin}_{\theta} \sum_{i=1}^n (y_i - \theta^T x_i)^2 = \|y - X\theta\|_2^2$$

Outline for today

- Review of maximum likelihood regression
- **Maximum a priori regression**
- Bias-variance tradeoff

Four levels for ML problems

1. Data & application
2. Model
3. Optimization problem
4. Optimization algorithm

X, y . Goal: predict y .

Linearly parameterized

$$\min_w ||Xw - y||_2^2$$

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

Four levels for ML problems

1. Data & application

X, y . Goal: predict y .

2. Model

Linearly parameterized

3. Optimization problem

$$\min_w ||Xw - y||_2^2 + \lambda ||w||_2^2$$

4. Optimization algorithm

$$\hat{w} = (X^T X + \lambda I)^{-1} X^T y$$

Justifying least squares

1. Define a probabilistic model for our data
2. Define what it means for a model to be “good” probabilistically (i.e., MLE)
3. Do some math to show that the “best” linear model probabilistically corresponds to the OLS estimate

Justifying ridge regression

1. Define a probabilistic model for our data
2. Define what it means for a model to be “good” probabilistically (i.e., MAP)
3. Do some math to show that the “best” linear model probabilistically corresponds to the OLS estimate

Maximum a priori

f_θ = Family of models, e.g., $f_\theta(x) = \theta^T x$

$$\hat{\theta} = \operatorname{argmax}_\theta \mathcal{L}(\theta; \mathcal{D})$$

MLE: $\mathcal{L}(\theta; \mathcal{D}) = p(\text{data} = \mathcal{D} \mid \text{model} = f_\theta)$

MAP: $\mathcal{L}(\theta; \mathcal{D}) = p(\text{model} = f_\theta \mid \text{data} = \mathcal{D})$

Maximum a priori

MAP: $\mathcal{L}(\theta; \mathcal{D}) = p(\text{model} = f_\theta \mid \text{data} = \mathcal{D})$

How to compute this quantity?

Maximum a priori derivation

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} p(\text{model} = f_{\theta} \mid \text{data} = \mathcal{D})$$

$$= \arg \max_{\theta} \frac{p(\text{data} = \mathcal{D} \mid \text{model} = f_{\theta}) \cdot (\cdot)}{(\cdot)}$$

Maximum a priori derivation

$$\begin{aligned}\hat{\theta}_{\text{MAP}} &= \arg \max_{\theta} p(\text{model} = f_{\theta} \mid \text{data} = \mathcal{D}) \\ &= \arg \max_{\theta} \frac{p(\text{data} = \mathcal{D} \mid \text{model} = f_{\theta}) \cdot p(\text{model} = f_{\theta})}{p(\text{data} = \mathcal{D})} \\ &= \arg \max_{\theta} p(\text{data} = \mathcal{D} \mid \text{model} = f_{\theta}) \cdot p(\text{model} = f_{\theta}) \\ &= \arg \max_{\theta} \log p(\text{data} = \mathcal{D} \mid \text{model} = f_{\theta}) + \log p(\text{model} = f_{\theta}) \\ &= \arg \min_{\theta} -\log p(\text{data} = \mathcal{D} \mid \text{model} = f_{\theta}) - \log p(\text{model} = f_{\theta})\end{aligned}$$

Maximum a priori vs MLE

$$\hat{\theta}_{\text{MLE}} = \arg \min_{\theta} -\log p(\text{data} = \mathcal{D} \mid \text{model} = f_{\theta})$$

$$\hat{\theta}_{\text{MAP}} = \arg \min_{\theta} -\log p(\text{data} = \mathcal{D} \mid \text{model} = f_{\theta}) - \log p(\text{model} = f_{\theta})$$



Same as MLE

Prior

Maximum a priori derivation

Assumption: $\theta^j \stackrel{iid}{\sim} \mathcal{N}(\mu_{\text{prior}}^j, \sigma_{\text{prior}}^2)$

$$\begin{aligned}\hat{\theta}_{\text{MAP}} &= \arg \min_{\theta} -\log p(\text{data} = \mathcal{D} \mid \text{model} = f_{\theta}) - \log p(\text{model} = f_{\theta}) \\ &= \arg \min_{\theta} - \left(\sum_{i=1}^n \log [p(y_i \mid x_i, \theta)] \right) - \log p(\theta) \\ &= \arg \min_{\theta} \frac{\sum_{i=1}^n (y_i - f_{\theta}(x_i))^2}{2\sigma^2} + \frac{\sum_{j=1}^d (\theta^j - \mu_{\text{prior}}^j)^2}{2\sigma_{\text{prior}}^2} \\ &= \arg \min_{\theta} \sum_{i=1}^n (y_i - f_{\theta}(x_i))^2 + \frac{\sigma^2}{\sigma_{\text{prior}}^2} \sum_{j=1}^d (\theta^j - \mu_{\text{prior}}^j)^2\end{aligned}$$

Maximum a priori derivation

$$= \arg \min_{\theta} \sum_{i=1}^n (y_i - f_{\theta}(x_i))^2 + \frac{\sigma^2}{\sigma_{\text{prior}}^2} \sum_{j=1}^d (\theta^j - \mu_{\text{prior}}^j)^2$$

assume $\mu_{\text{prior}}^j = 0$

let $\lambda = \frac{\sigma^2}{\sigma_{\text{prior}}^2}$

$$\arg \min_{\theta} \sum_{i=1}^n (y_i - f_{\theta}(x_i))^2 + \lambda \sum_{j=1}^d (\theta^j)^2$$

Ridge regression

Outline for today

- Review of maximum likelihood estimation
- Maximum a priori
- **Bias-variance tradeoff**

Bias-variance tradeoff

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$$

View x_i as sampled from r.v. X

Goal: We want to find the model family with the lowest *expected* error

Bias-variance tradeoff

Goal: We want to find the model family with the lowest *expected* error

$$\epsilon(\vec{x}; f) = \mathbb{E} [(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2] = \mathbb{E}_{\vec{x}, y, \mathcal{D}} [(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2]$$

$f_\theta^*(\vec{x}; \mathcal{D})$: best model from the family on the given dataset

Bias-variance tradeoff

Goal: We want to find the model family with the lowest *expected* error

$$\epsilon(\vec{x}; f) = \mathbb{E}_{\vec{x}, y, \mathcal{D}} [(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2]$$

Let's decompose this into parts

Bias-variance tradeoff

Goal: We want to find the model family with the lowest *expected* error

$$\epsilon(\vec{x}; f) = \mathbb{E}_{\vec{x}, y, \mathcal{D}} [(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2]$$

Let's decompose this into parts

$$= (\mathbb{E}[f_\theta^*(\vec{x}; \mathcal{D})] - \mathbb{E}[y])^2 + \text{Var}(f_\theta^*(\vec{x}; \mathcal{D})) + \text{Var}(z)$$

The equation is decomposed into three terms by horizontal lines and arrows pointing downwards. The first term, $(\mathbb{E}[f_\theta^*(\vec{x}; \mathcal{D})] - \mathbb{E}[y])^2$, is labeled 'bias^2'. The second term, $\text{Var}(f_\theta^*(\vec{x}; \mathcal{D}))$, is labeled 'Variance'. The third term, $\text{Var}(z)$, is labeled 'Irreducible error'.

bias²

Variance Irreducible
error

Bias-variance tradeoff

Goal: We want to find the model family with the lowest *expected* error

$$\epsilon(\vec{x}; f) = \mathbb{E}_{\vec{x}, y, \mathcal{D}} [(f_{\theta}^*(\vec{x}; \mathcal{D}) - y)^2]$$

Start by analyzing mean & var of y

$$\mathbb{E}[Y] = \mathbb{E}[f(\vec{x}) + z] = f(\vec{x}) + \mathbb{E}[z] = f(\vec{x})$$

$$\text{Var}(Y) = \text{Var}(f(\vec{x}) + z) = \text{Var}(Z)$$

Bias-variance tradeoff

Goal: We want to find the model family with the lowest *expected* error

$$\epsilon(\vec{x}; f) = \mathbb{E}_{\vec{x}, y, \mathcal{D}} [(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2]$$

Add one more lemma (for any RV)

$$\text{Var}(X) = \mathbb{E} [(X - \mathbb{E}(X))^2] = \mathbb{E} [X^2] - \mathbb{E} [X]^2$$

$$\implies \mathbb{E} [X^2] = \text{Var}(X) + \mathbb{E} [X]^2$$

Bias-variance tradeoff

$$\begin{aligned}\epsilon(\vec{x}; f) &= \mathbb{E}_{\vec{x}, y, \mathcal{D}} [(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2] \\&= \mathbb{E}(f_\theta^*(\vec{x}; \mathcal{D})^2) + \mathbb{E}(y^2) - 2\mathbb{E}[f_\theta^*(\vec{x}; \mathcal{D}) \cdot y] \\&= (\text{Var}(f_\theta^*(\vec{x}; \mathcal{D})) + \mathbb{E}[f_\theta^*(\vec{x}; \mathcal{D})]^2) \\&\quad + (\text{Var}(Y) + \mathbb{E}[y]^2) - 2\mathbb{E}[f_\theta^*(\vec{x}; \mathcal{D})] \cdot \mathbb{E}[y] \\&= (\mathbb{E}[f_\theta^*(\vec{x}; \mathcal{D})]^2 - 2\mathbb{E}[f_\theta^*(\vec{x}; \mathcal{D})] \cdot \mathbb{E}[y] + \mathbb{E}[y]^2) \\&\quad + \text{Var}(f_\theta^*(\vec{x}; \mathcal{D})) + \text{Var}(z) \\&= (\mathbb{E}[f_\theta^*(\vec{x}; \mathcal{D})] - \mathbb{E}[y])^2 + \text{Var}(f_\theta^*(\vec{x}; \mathcal{D})) + \text{Var}(z)\end{aligned}$$

Facts

$$\mathbb{E}[X^2] = \text{Var}(X) + \mathbb{E}[X]^2$$

$$\text{Var}(Y) = \text{Var}(z)$$

$$\mathbb{E}(Y) = f(\vec{x})$$

Bias-variance tradeoff

$$\epsilon(\vec{x}; f) = \mathbb{E}_{\vec{x}, y, \mathcal{D}} [(f_\theta^*(\vec{x}; \mathcal{D}) - y)^2]$$
$$= \underbrace{(\mathbb{E} [f_\theta^*(\vec{x}; \mathcal{D})] - \mathbb{E} [y])^2}_{\text{bias}^2} + \underbrace{\text{Var}(f_\theta^*(\vec{x}; \mathcal{D}))}_{\text{Variance}} + \underbrace{\text{Var}(z)}_{\text{Irreducible error}}$$