# CS189: Intro to Machine Learning
# Summer 2018

Lecture 4: Maximum a posteriori 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 posteriori
- Bias-variance tradeoff

# Outline for today

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

# Four levels for ML problems

| | |
|---|---|
| 1. Data & application | X, y. Goal: predict y. |
| 2. Model | Linearly parameterized |
| 3. Optimization problem | $\min_w \|\|Xw - y\|\|_2^2$ |
| 4. Optimization algorithm | $\hat{w} = (X^T X)^{-1} X^T y$ |

# 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

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

Therefore:

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

# Maximum Likelihood Estimation

$f_\theta$ = *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

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

# MLE justifies OLS

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

Linear regression model:

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

# Outline for today

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

# Four levels for ML problems

| | |
|---|---|
| 1. Data & application | X, y. Goal: predict y. |
| 2. Model | Linearly parameterized |
| 3. Optimization problem | $\min_w \|\|Xw - y\|\|_2^2$ |
| 4. Optimization algorithm | $\hat{w} = (X^T X)^{-1} X^T y$ |

# 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 posteriori

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

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

**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 posteriori

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

How to compute this quantity?

# Maximum a posteriori derivation

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

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

# Maximum a posteriori derivation

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

# Maximum a posteriori vs MLE

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

$$\hat{\theta}_{\mathrm{MAP}} = \arg\min_\theta \underbrace{- \log p(\mathrm{data} = \mathcal{D} \mid \mathrm{model} = f_\theta)}_{\text{Same as MLE}} \underbrace{- \log p(\mathrm{model} = f_\theta)}_{\textit{Prior}}$$

# Maximum a posteriori derivation

**Assumption:** $\theta^j \overset{iid}{\sim} \mathcal{N}(\mu^j_{\text{prior}}, \sigma^2_{\text{prior}})$

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

$$= \arg\min_\theta - \left( \sum_{i=1}^{n} \log \left[ p(y_i \mid x_i, \theta) \right] \right) - \log p(\theta)$$

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

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

# Maximum a posteriori derivation

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

$$\text{assume } \mu^j_{\text{prior}} = 0 \qquad \text{let } \lambda = \frac{\sigma^2}{\sigma^2_{\text{prior}}}$$

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

**Ridge regression**

# Outline for today

- Review of maximum likelihood estimation
- Maximum a posteriori
- **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

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

$f_\theta^*(\vec{x}; \mathcal{D})$: 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

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

**Let's decompose this into parts**

# Bias-variance tradeoff

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

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

**Let's decompose this into parts**

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

# Bias-variance tradeoff

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

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

**Start by analyzing mean & var of y**

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

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

# Bias-variance tradeoff

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

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

**Add one more lemma (for any RV)**

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

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

# Bias-variance tradeoff

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

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

$$= \left(\mathrm{Var}(f_\theta^*(\vec{x};\mathcal{D})) + \mathbb{E}\left[f_\theta^*(\vec{x};\mathcal{D})\right]^2\right)$$

$$+ \left(\mathrm{Var}(Y) + \mathbb{E}\left[y\right]^2\right) - 2\mathbb{E}\left[f_\theta^*(\vec{x};\mathcal{D})\right] \cdot \mathbb{E}\left[y\right]$$

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

$$+ \mathrm{Var}(f_\theta^*(\vec{x};\mathcal{D})) + \mathrm{Var}(z)$$

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

## Facts

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

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

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

# Bias-variance tradeoff

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

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