Estimating Probabilities from Data

Cornell CS 3/5780 · Spring 2026

1. Motivation: Bayes Optimal Classifier

- **Recall**: Bayes Optimal classifier predicts $\arg\max_y P(y|\mathbf{x})$
- **Goal**: Can we estimate $P(X, Y)$ directly from training data?
- **Two approaches**:
  - Generative learning: Estimate $P(X, Y) = P(X|Y)P(Y)$
  - Discriminative learning: Estimate $P(Y|X)$ directly
- How can we estimate probability distributions from samples?
- **Example**: Tossing a possibly biased coin.

2. Maximum Likelihood Estimation (MLE)

- **Two-step procedure**:
  1. Make assumption about distribution of data $P(D;\theta)$
  2. Set parameters θ to maximize likelihood of observed data
- **MLE Principle**: Find $\hat{\theta}$ to maximize likelihood

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D;\theta)$$

- **Example**: binomial distribution models n independent Bernoulli trials with probability θ

$$P(D;\theta) = \binom{n_H + n_T}{n_H} \theta^{n_H} (1-\theta)^{n_T}$$

3. MLE Derivation

- **General procedure**: To solve for $\hat{\theta}_{MLE}$
  1. Plug data into distribution and take logarithm: $\log P(D; \theta)$
  2. Take the derivative and set it to zero
- **Question**: What is the MLE derivation for the coin toss with a binomial distribution?

- **Pros**: If n is large and model is correct, finds *true* parameters
- **Cons**: Can overfit when n is small and can be wrong if model is incorrect

4. Incorporating Prior Knowledge

- **Idea**: Add imaginary data that mirrors our prior knowledge
  - Example: m_H imaginary heads and m_T imaginary tails

$$\hat{\theta} = \frac{n_H + m_H}{n_H + n_T + m_H + m_T}$$

- **Bayesian Formalization**: Model θ as a *random variable* with *prior* distribution $P(\theta)$
- **Bayes Rule**:

$$P(\theta \mid D) = \frac{P(D \mid \theta)P(\theta)}{P(D)}$$

- **Components**:
  - $P(\theta)$: *prior* distribution (before seeing data)
  - $P(D \mid \theta)$: *likelihood* of data
  - $P(\theta \mid D)$: *posterior* distribution (after seeing data)

5. Maximum a Posteriori (MAP)

- **Two-step procedure**:
  1. Make assumption about distribution of data *and the distribution of* θ
  2. Set parameters to maximize likelihood of observed data *and parameters*
- **MAP Principle**: Choose most likely θ given data *and prior distribution*

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\, P(\theta \mid D)$$

$$= \underset{\theta}{\operatorname{argmax}} \log P(D|\theta) + \log P(\theta)$$

- **Example**: Beta distribution as a coin prior

$$P(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}$$

- **Question**: What is the MAP derivation for the coin toss with a binomial distribution and a beta prior? How does it relate to "imaginary" data?

6. MLE and MAP Summary

Given training data D, parameters θ, test point x_t:

MLE:

- Prediction: $P(y \mid x_t; \theta)$
- Learning: $\theta = \mathrm{argmax}_\theta \, P(D; \theta)$
- θ is a model parameter
- Works if n is large enough and model is correct

MAP:

- Prediction: $P(y \mid x_t, \theta)$
- Learning: $\theta = \mathrm{argmax}_\theta \, P(\theta \mid D) \propto P(D \mid \theta)P(\theta)$
- θ is a random variable
- $\log[P(\theta)]$ penalizes deviating from prior belief
- Can work for smaller n *if* the prior is correct (*and* the model)

Convergence: As $n \to \infty$, $\hat{\theta}_{MAP} \to \hat{\theta}_{MLE}$

7. Estimating Distributions for ML

- Training data: $D = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ drawn i.i.d. from $P(X, Y)$
- Joint distribution:
  - $\hat{P}(\mathbf{x}, y) =$
- Marginal distributions:
  - $\hat{P}(y) =$
  - $\hat{P}(\mathbf{x}) =$
- Conditional distributions:
  - $\hat{P}(\mathbf{x}|y) =$
  - $\hat{P}(y|\mathbf{x}) =$