Probabilistic Graphical Models & Probabilistic AI

Ben Lengerich

Lecture 2: Statistics Review

January 23, 2025

Reading: See course homepage

Logistics Review

- Class webpage: lengerichlab.github.io/pgm-spring-2025
- Lecture scribe sign-up sheet
- Readings: **Canvas**
- Class Announcements: **Canvas**
- Assignment Submissions: **Canvas**
- Instructor: **Ben Lengerich**
  - Office Hours: Thursday 2:30-3:30pm, 7278 Medical Sciences Center
  - Email: lengerich@wisc.edu
- TA: **Chenyang Jiang**
  - Office Hours: Monday 11am-12pm, 1219 Medical Sciences Center
  - Email: cjiang77@wisc.edu

Homework 1

- Released, due next Friday at midnight.
  - PDF and Latex solution template (.tex) available on website.
- Submit via **Canvas**.
- Most preferred format:
  - PDF with your solution written in the provided solution box using Latex.
- Questions – Ask early and often

```
\begin{solution}
    Write your solution here. For multiple choice
    questions, only the letter answer is required.
  \begin{parts}
    \part Solution for (a)
    \part Solution for (b)
    \part Solution for (c)
    \part Solution for (d)
  \end{parts}
\end{solution}
```

Answer: Write your solution here.

(a) Solution for (a)

(b) Solution for (b)

(c) Solution for (c)

(d) Solution for (d)

Questions about Course Logistics?

Statistics Review

Today

- Probability Basics
- Estimation Methods
- Linear Regression
- Optimization

Probability Basics

Probability Basics: Definitions

- Random Variables:
  - Discrete: Values from a countable set (e.g. a coin flip)
  - Continuous: Values from an interval (e.g. a height)
- PMF and PDF:
  - **P**robability **M**ass **F**unction: P(X=x) for discrete X.
  - **P**robability **D**ensity **F**unction: f(x) for continuous X.

Key Distributions

- Bernoulli Distribution:
  - $P(X = x) = \theta^x (1 - \theta)^{1-x}, x \in \{0,1\}$
  - Example: a fair coin flip ($\theta = 0.5$)
- Gaussian Distribution:
  - $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$
  - **“Normal”** because of Central Limit Theorem
  - “Standard Normal” when $\mu = 0, \sigma = 1$

Central Limit Theorem

- Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with mean μ and variance σ^2.
- Define the sample mean:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Then, as $n \to \infty$:

$$\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \to N(0,1)$$

Joint, Marginal, and Conditional Probabilities

- Joint: $P(A, B)$, probability of two events occurring together.
- Marginal: $P(A) = \sum_B P(A, B)$, sum of joint probabilities over one variable.
- Conditional: $P(A|B) = \frac{P(A,B)}{P(B)}$, probability of A given B.

Expectation and Variance

- Expectation:
  - Discrete: $E[X] = \sum_x xP(X = x)$
  - Continuous: $E[X] = \int xf(x)dx$
- Variance: $Var(X) = E[(X - E[X])^2]$
  - Equivalent: $Var(X) = E[X^2] - E[X]^2$

Linearity of Expectation

- Property:
  - $E[aX + b] = aE[X] + b$
- Multiple Variables:
  - $E[X_1 + X_2] = E[X_1] + E[X_2]$

Expectation of Functions

- Formula:
  - $E[g(X)] = \sum_x g(x)P(X = x)$ (discrete)
  - $E[g(X)] = \int_x g(x)f(x)dx$ (continuous)

- Example (Discrete):
  - $X \sim \text{Bernoulli}(\theta), g(X) = X^2$:
  - $E[g(X)] = 1^2\theta + 0^2(1-\theta) = \theta$
- Example (Continuous):
  - $X \sim \text{Uniform}(0,1), g(X) = X^2$:
  - $E[g(X)] = \int_0^1 x^2 dx = \frac{1}{3}$.

Variance of Functions

- Definition:
  - $Var\big(g(X)\big) = E[(g(X) - E[g(X)])^2]$
  - Equivalent: $Var\big(g(X)\big) = E[g(X)^2] - (E[g(X)])^2$

Covariance and Correlation

- Covariance:
  - $Cov(X,Y) = E[(X - E[X])(Y - E[Y])]$
- Properties:
  - $Cov(X,X) = Var(X)$
  - If X,Y are independent: $Cov(X,Y) = 0$.
- Correlation:
  - $\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$
  - $\rho = 1$: Perfect positive linear relationship.
  - $\rho = 0$: No linear relationship.
  - $\rho = -1$: Perfect negative linear relationship.

Bayes' Rule

- $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
- Example: Medical test:
  - $P(disease|positive\ test) = \frac{P(positive\ test\ |disease)\ P(disease)}{P(positive\ test)}$

Estimation Methods

Introduction to Estimation

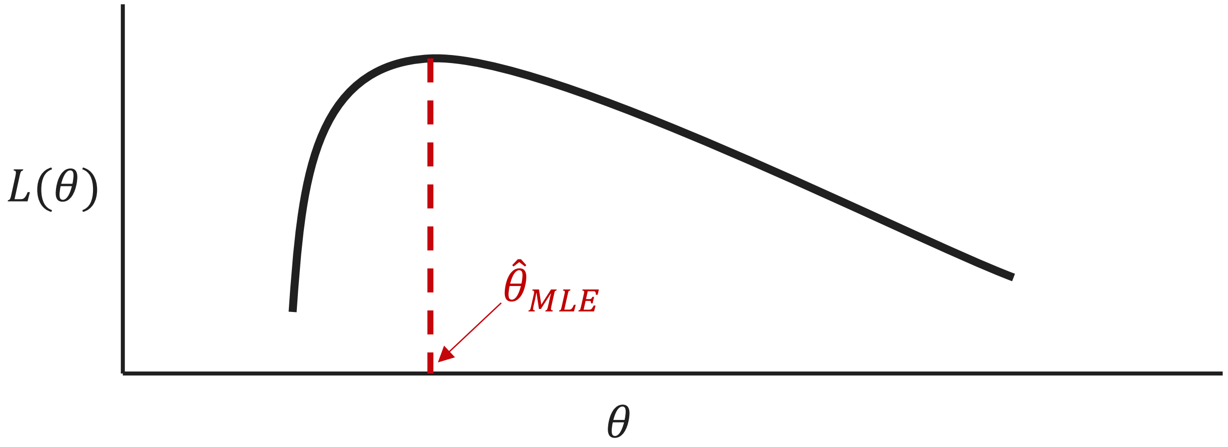
- **Goal of Estimation:**
  - Infer unknown parameters θ from observed data.
- **Types of Estimation:**
  - Point Estimation: Single value (e.g., MLE).
  - Interval Estimation: Range of plausible values (e.g., confidence intervals).
- **Common Methods:**
  - Maximum Likelihood Estimation (MLE)
  - Maximum A Posteriori (MAP)
  - Method of Moments

Maximum Likelihood Estimation (MLE)

- **Definition:**
  - Find $\hat{\boldsymbol{\theta}}$ that maximizes the likelihood of observing the given data.

$$\hat{\boldsymbol{\theta}} = \text{argmax}_\theta L(\theta) \text{ where } L(\theta) = P(\text{data}|\theta).$$

- **Interpretation:**
  - L(θ): Probability of the observed data given θ.
  - MLE chooses the parameter that makes the data most "likely."



Maximum Likelihood Estimation (MLE)

- **Example:**
  - Dataset: X={1,0,1,1,0},
  - Bernoulli distribution with $P(X = 1|\theta) = \theta$:

$$L(\theta) = \prod_i \theta^{x_i} (1-\theta)^{1-x_i}$$

  - Typically solved by maximizing the log-likelihood.

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} (x_i \log \theta \ + (1 - x_i) \log(1-\theta))$$

  - Derivative:

$$\frac{d\ell}{d\theta} = \frac{k}{\theta} - \frac{n-k}{1-\theta}$$

where $k = \sum x_i$

  - Solution:

$$\hat{\boldsymbol{\theta}} = \frac{k}{n}$$

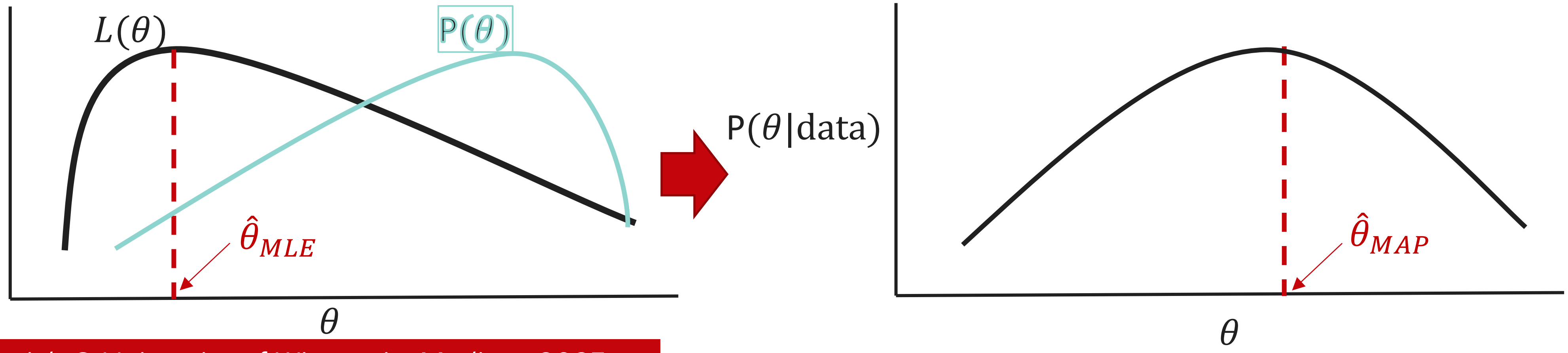
Maximum Likelihood Estimation (MLE)

- The MLE:
  - does not always exist.
  - is not necessarily unique.
  - is not necessarily admissible.

Maximum A Posteriori (MAP) Estimation

- Find
$$\hat{\theta}_{MAP} = argmax_{\theta} P(\theta|\text{data}) \propto argmax_{\theta} P(\text{data}|\theta) P(\theta)$$
- $P(\text{data}|\theta)$: Likelihood
- $P(\theta)$: Prior belief about θ

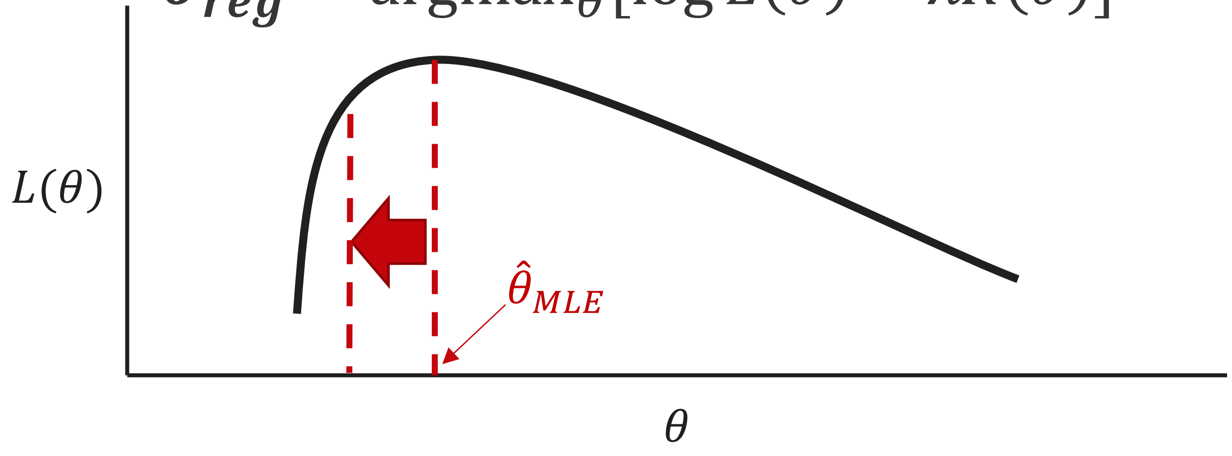
- MLE ignores $P(\theta)$
- MAP incorporates prior information.



Regularization is MAP

- **MLE with Regularization:**
  - Adds a penalty to avoid overfitting

$$\widehat{\boldsymbol{\theta}_{\boldsymbol{reg}}} = \text{argmax}_\theta[\log L(\theta) - \lambda R(\theta)]$$



- **MAP as Penalized MLE:**
  - Let $P(\theta) \propto e^{-\lambda R(\theta)}$. Then

$$\widehat{\boldsymbol{\theta}}_{MAP} = argmax_\theta[\log L(\theta) + \log P(\theta)] = \hat{\theta}_{reg}$$

Method of Moments

- **Definition:**
  - Match sample moments to theoretical moments ($E[X^n]$) to estimate parameters.
- **Example:**
  - Bernoulli:
    - $E[X] = \theta$, estimate $\widehat{\boldsymbol{\theta}} = \bar{X}$.
  - Gaussian:
    - $E[X] = \mu$.
    - $Var(X) = \sigma^2$

Linear Regression

Introduction to Linear Regression

- **Model Definition:**
  - $y = X\beta + \epsilon$, where
  - y: Response variable (dependent variable).
  - X: Design matrix (independent variables or features).
  - β: Coefficients (parameters to estimate).
  - ϵ: Error term (often assumed to be $N(0, \sigma^2)$).
- **Goal:**
  - Estimate β.

Linear Regression Evaluation Metrics

- **Coefficient of Determination (R^2):**
  - $R^2 = 1 - \frac{SS_{residual}}{SS_{total}}$
  - Measures the proportion of variance explained by the model.
- **Mean Squared Error (MSE):**
  - $MSE = \frac{1}{n}\sum(y_i - \hat{y}_i)^2$
- **Mean Absolute Error (MAE):**
  - $MSE = \frac{1}{n}\sum\|y_i - \hat{y}_i\|$

Ordinary Least Squares (OLS)

- **Objective:**
  - Minimize the sum of squared residuals:
  - $\hat{\boldsymbol{\beta}}_{OLS} = argmin_{\beta} \|y - X\beta\|^2$
- Residuals:
  - $e_i = y_i - \hat{y}_i$
- **Solution:**
  - $\hat{\boldsymbol{\beta}}_{OLS} = (X^T X)^{-1} X^T Y$

Regularization in Linear Regression (MAP)

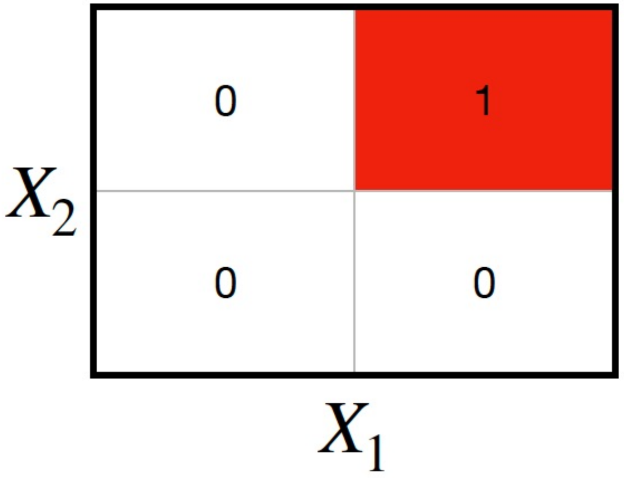
- **Ridge Regression (L2 Regularization):**
  - Adds an L2 penalty:
    - $\widehat{\boldsymbol{\beta}}_{ridge} = argmin_{\beta}\|y - X\beta\|^2 + \lambda\|\beta\|^2$
  - Equivalent MAP interpretation:
    - Prior on coefficients: $\beta \sim N(0, \frac{\sigma^2}{\lambda})$
    - MAP estimate maximizes: $P(\beta|y) \propto P(y|\beta)P(\beta)$
    - Penalty comes from the Gaussian prior.
- **Lasso Regression (L1 Regularization):**
  - Adds an L1 penalty:
    - $\widehat{\boldsymbol{\beta}}_{lasso} = argmin_{\beta}\|y - X\beta\|^2 + \lambda\|\beta\|_1$
  - Equivalent to $\beta \sim \text{Laplace}(0, \frac{\sigma}{\lambda})$

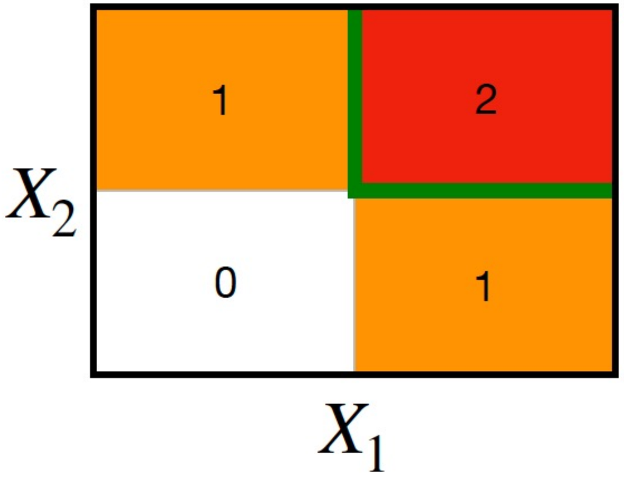
Extensions of Linear Regression

- **Polynomial Regression:**
  - Add polynomial terms:
  - $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ \ldots$
- **Generalized Linear Models:**
  - Extend to non-normal distributions by a link function:
  - $g(E[Y]) = X\beta$
- **Interaction Terms:**
  - Include interactions between predictors:
  - $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ + \beta_3 x_1 x_2$

A word of warning on interpreting interactions...

- Suppose we have data from:
  $$Y = AND(X_1, X_2)$$
- with Boolean X. Let’s fit an additive model (no interactions):
- $\hat{Y} = f_0 + f_1(X_1) + f_2(X_2)$
- How well can we fit the data?





Optimization

Convexity

Convex function

$$\forall\lambda \in [0,1], \qquad f(\lambda\,\mathbf{x} + (1-\lambda)\,\boldsymbol{y}) \le \lambda\, f(\mathbf{x}) + (1-\lambda)\, f(\boldsymbol{y})$$

Strictly convex function

$$\forall\lambda \in]0,1[, \qquad f(\lambda\,\mathbf{x} + (1-\lambda)\,\boldsymbol{y}) < \lambda\, f(\mathbf{x}) + (1-\lambda)\, f(\boldsymbol{y})$$

Strongly convex function

$$\exists\mu > 0, \text{ s.t. } \quad \mathbf{x} \mapsto f(\mathbf{x}) - \mu\|\mathbf{x}\|^2 \quad \text{is convex}$$

Equivalently:

$$\forall\lambda \in [0,1], \quad f(\lambda\,\mathbf{x}+(1-\lambda)\,\boldsymbol{y}) \le \lambda\, f(\mathbf{x})+(1-\lambda)\, f(\boldsymbol{y})-\mu\,\lambda(1-\lambda)\|\mathbf{x}-\boldsymbol{y}\|^2$$

The largest possible μ is called the strong convexity constant.

Convexity Aids Optimization

If f is convex and differentiable at $\mathbf{x}$ then

$$f(\boldsymbol{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\boldsymbol{y} - \mathbf{x})$$

Convex function
All local minima are global minima.

Strictly convex function
If there is a local minimum, then it is unique and global.

Strongly convex function
There exists a unique local minimum which is also global.

But...

Convexity is Overrated

- **Using a suitable architecture (even if it leads to non-convex loss functions) is more important than insisting on convexity (particularly if it restricts us to unsuitable architectures)**
  - e.g.: Shallow (convex) classifiers versus Deep (non-convex) classifiers
- **Even for shallow/convex architecture, such as SVM, using non-convex loss functions actually improves the accuracy and speed**
  - See "trading convexity for efficiency" by Collobert, Bottou, and Weston, ICML 2006 (best paper award)

- Yann LeCun, "Who's afraid of Non-convex loss functions?" – **2007**

Questions?

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON