Bayesian Estimation

Outline

- Review of maximum likelihood estimation
- Bayesian estimation
- Examples in machine learning

Maximum Likelihood Estimation

- A *probability model* explains how data are generated, given a parameter $\boldsymbol{\theta}$

$$\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n \overset{iid}{\sim} p(\boldsymbol{z}; \boldsymbol{\theta})$$

  where $p(\boldsymbol{z}; \boldsymbol{\theta})$ is a pdf/pmf

- Given a realization of data $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$, the *likelihood function* is

$$L(\boldsymbol{\theta}) := \prod_{i=1}^{n} p(\boldsymbol{z}_i; \boldsymbol{\theta})$$

  and the *log-likelihood* is

$$\ell(\boldsymbol{\theta}) := \log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\boldsymbol{z}_i; \boldsymbol{\theta}).$$

- A *maximum likelihood estimate* (MLE) is any

$$\hat{\theta} \in \arg\max_{\theta} \ell(\theta)$$

MLE Examples

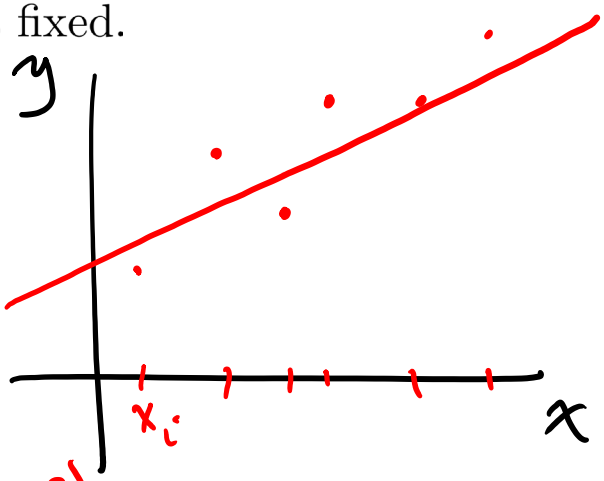
- LDA
- Naïve Bayes
- Logistic regression
- Linear regression ← today
- Gaussian mixture models

Maximum Likelihood Linear Regression

- Training data: $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$. View $\boldsymbol{x}_i$ as fixed.
- No offset, $\boldsymbol{\theta} = \boldsymbol{w}$
- Gaussian likelihood:

$$y_i = w^T x_i + \epsilon_i$$

where $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_e^2)$, σ_e^2 known.



$$\Leftrightarrow \quad y_i \sim N(w^T x_i, \sigma_e^2)$$

- Likelihood

$$L(w) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left\{ -\frac{1}{2\sigma_e^2} (y_i - w^T x_i)^2 \right\}$$

- Log-likelihood

$$-\frac{1}{2\sigma_e^2} \sum_{i=1}^{n} (y_i - w^T x_i)^2 + \text{const.}$$

MLE $\Leftrightarrow$ OLS

Bayesian Estimation

- Maximum likelihood assumes that $\boldsymbol{\theta}$ is nonrandom
- In Bayesian estimation, $\boldsymbol{\theta}$ is viewed as random
- Our uncertainty about $\boldsymbol{\theta}$ depends on whether we have observed data or not
- Let $\underline{\boldsymbol{Z}} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)$ denote all observed quantities
- Let $p(\boldsymbol{\theta})$ denote the distribution of $\boldsymbol{\theta}$ *before* observing a realization $\underline{\boldsymbol{z}}$ of $\underline{\boldsymbol{Z}}$.
- Let $p(\boldsymbol{\theta} \mid \underline{\boldsymbol{z}})$ denote the distribution of $\boldsymbol{\theta}$ *after* observing a realization $\underline{\boldsymbol{z}}$ of $\underline{\boldsymbol{Z}}$.
- These are related by Bayes rule:

$$p(\theta \mid \underline{z}) = \frac{p(\underline{z} \mid \theta)\, p(\theta)}{p(\underline{z})}$$

user-specified

$(\underline{Z}, \theta)$ jointly distributed

prior distribution

posterior distribution

Bayesian Estimation

- A parameter estimate can be obtained from the posterior in multiple ways:
- Posterior mean: $\hat{\theta}(\underline{z}) = \mathbb{E}[\theta \mid \underline{Z} = \underline{z}] = \int \theta\, p(\theta \mid \underline{z})\, d\theta$

- Maximum a posteriori (MAP) estimation:

$$\hat{\theta}(\underline{z}) = \arg\max_{\theta} \log p(\theta \mid \underline{z})$$

$$= \arg\max_{\theta} \log p(\underline{z} \mid \theta) + \log p(\theta)$$

- In addition to parameter estimates, the main advantage of Bayesian methods over frequentist methods is they also provide natural confidence intervals.

Bayesian Linear Regression

- Training data: $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$. View $\boldsymbol{x}_i$ as fixed.
- No offset, $\boldsymbol{\theta} = \boldsymbol{w}$
- Gaussian likelihood:

$$y_i = \boldsymbol{w}^T \boldsymbol{x}_i + \epsilon_i, \quad \Longleftrightarrow \quad y_i \sim N(w^T x_i, \sigma_e^2)$$

  where $\epsilon_i \sim \mathcal{N}(0, \sigma_e^2)$

- Prior:

$$w \sim N(0, \sigma_w^2 \mathbf{I}) \Longleftrightarrow p(w) = (2\pi\sigma_w^2)^{-\frac{d}{2}} \exp\left\{-\frac{\|w\|^2}{2\sigma_w^2}\right\}$$

$$\Longleftrightarrow \quad w = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}, \quad w_i \overset{iid}{\sim} N(0, \sigma_w^2)$$

- σ_e^2, σ_w^2 are hyperparameters

Bayesian Linear Regression

- Joint distribution:

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}, \quad X = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}$$

$(d \times n)$

$$\begin{bmatrix} y \\ w \end{bmatrix} = \left[\begin{array}{c:c} I & X^T \\ \hdashline 0 & I \end{array}\right] \begin{bmatrix} \epsilon \\ w \end{bmatrix}$$

$(n+d) \times 1$

Poll

Which of the following best describes the posterior distribution of $\boldsymbol{w}$ given $\boldsymbol{y}$:

(A) Gaussian

(B) Mixture of Gaussians

(C) Gaussian with low rank covariance

(D) None of the above

, i.e. the distribution

Bayesian Linear Regression

- Posterior:

$$w|y \sim \mathcal{N}(\mu(y), \Sigma(y))$$

where

$$\hat{w} = \boldsymbol{\mu}(\boldsymbol{y}) = \left(\boldsymbol{X}\boldsymbol{X}^T + \frac{\sigma_e^2}{\sigma_w^2}\boldsymbol{I}\right)^{-1}\boldsymbol{X}\boldsymbol{y}$$

$$\boldsymbol{\Sigma}(\boldsymbol{y}) = \left(\frac{1}{\sigma_e^2}\boldsymbol{X}\boldsymbol{X}^T + \frac{1}{\sigma_w^2}\boldsymbol{I}\right)^{-1}$$

and

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_n \end{bmatrix} \in \mathbb{R}^{d \times n}$$

MAP Estimate

- Log-posterior

$$\log p(w|y) = \log \frac{p(y|w)\,p(w)}{p(y)}$$

$$= \log p(y|w) + \log p(w) + \text{const.}$$

$$= -\frac{1}{2\sigma_e^2}\sum_{i=1}^{n}(y_i - w^T x_i)^2 - \frac{1}{2\sigma_w^2}\|w\|^2 + \text{const}$$

MAP: $\max_w p(w|y) \iff \min_w \frac{1}{2\sigma_e^2}\sum_{i=1}^{n}(y_i - w^T x_i)^2 + \frac{1}{2\sigma_e^2}\|w\|^2$

- Compare with ridge regression

$$\min_w \frac{1}{n}\sum_{i=1}^{n}(y_i - w^T x_i)^2 + \lambda\|w\|^2, \quad \lambda = \frac{\sigma_e^2}{n\sigma_w^2}$$

Bayesian Linear Regression

- Linear transformation of a MVG is another MVG
- In particular, if $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\boldsymbol{A}$ is a matrix and $\boldsymbol{b}$ a column vector of appropriate dimensions, then

$$\boldsymbol{AZ} + \boldsymbol{b} \sim \mathcal{N}(\boldsymbol{A\mu} + \boldsymbol{b}, \boldsymbol{A\Sigma A}^T)$$

- If $\boldsymbol{x}$ is a test point, then the predicted output

$$\widehat{f}(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{w}$$

- Thus, the predicted output is a random variable with distribution

$$\widehat{f}(\boldsymbol{x}) \mid \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{x}^T\boldsymbol{\mu}(\boldsymbol{y}), \boldsymbol{x}^T\boldsymbol{\Sigma}(\boldsymbol{y})\boldsymbol{x})$$

$$\sim \mathcal{N}\left(\underbrace{\boldsymbol{x}^T\left(\boldsymbol{X}\boldsymbol{X}^T + \frac{\sigma_\varepsilon^2}{\sigma_w^2}\boldsymbol{I}\right)^{-1}\boldsymbol{X}\boldsymbol{y}}_{\mu_x(y)}, \underbrace{\boldsymbol{x}^T\left(\boldsymbol{X}\boldsymbol{X}^T + \frac{\sigma_\varepsilon^2}{\sigma_w^2}\boldsymbol{I}\right)^{-1}\boldsymbol{x}}_{\sigma_x^2(y)}\right)$$

Bayesian Linear Regression

- Using matrix identities (e.g., the matrix inversion lemma), it can be shown that

$$\widehat{f}(\boldsymbol{x}) \mid \boldsymbol{y} \sim \mathcal{N}(\mu_{\boldsymbol{x}}(\boldsymbol{y}), \sigma^2_{\boldsymbol{x}}(\boldsymbol{y}))$$

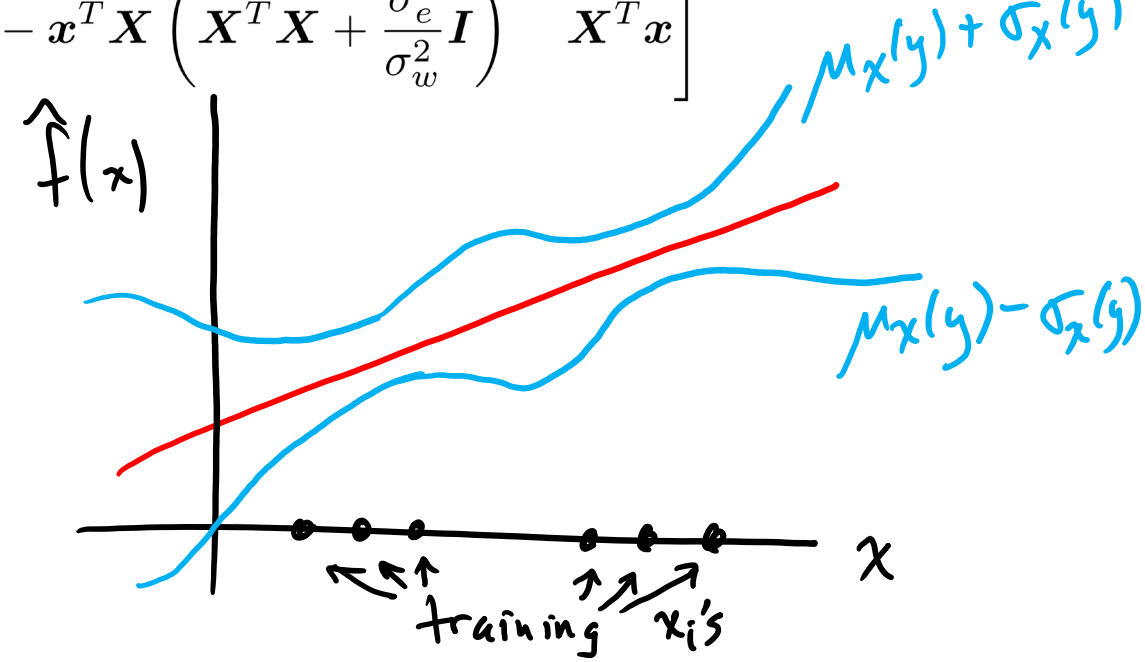
where

$$\mu_{\boldsymbol{x}}(\boldsymbol{y}) = \sigma_w^2 \boldsymbol{x}^T \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{X} + \frac{\sigma_e^2}{\sigma_w^2} \boldsymbol{I} \right)^{-1} \boldsymbol{y}$$

$$\sigma^2_{\boldsymbol{x}}(\boldsymbol{y}) = \sigma_w^2 \left[\boldsymbol{x}^T \boldsymbol{x} - \boldsymbol{x}^T \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{X} + \frac{\sigma_e^2}{\sigma_w^2} \boldsymbol{I} \right)^{-1} \boldsymbol{X}^T \boldsymbol{x} \right]$$

- Conclusion:

You can kernelize this method

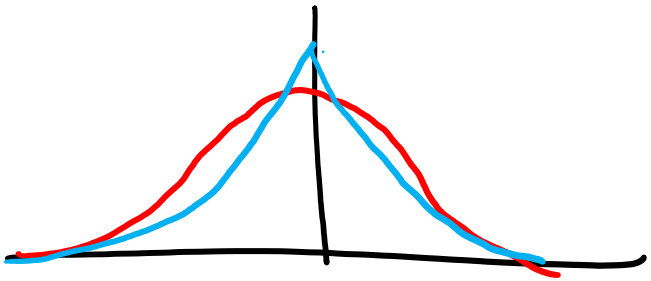


Laplacian Prior

- Training data: $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$. View $\boldsymbol{x}_i$ as fixed.
- No offset, $\boldsymbol{\theta} = \boldsymbol{w}$
- Gaussian likelihood:

$$y_i = \boldsymbol{w}^T \boldsymbol{x}_i + \epsilon_i,$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma_e^2)$



- Prior:

$$w = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}, \quad w_i \overset{iid}{\sim} \text{Laplacian}(\beta)$$

$$p(w_j) = \frac{\beta}{2} \exp\left(-\beta |w_j|\right)$$

MAP Estimate

- Log-posterior

$$\log p(w|y) = \log p(y|w) + \log p(w) + C$$

$$= -\frac{1}{2\sigma_e^2}\sum_{i=1}^{n}(y_i - w^T x_i)^2 - \beta\sum_{j=1}^{d}|w_j| + C$$

$$\text{MAP} \Leftrightarrow \min_w \frac{1}{2\sigma_e^2}\sum_{i=1}^{n}(y_i - w^T x_i)^2 + \beta\|w\|_1$$

- Compare with lasso regression

$$\min_w \frac{1}{n}\sum_{i=1}^{n}(y_i - w^T x_i)^2 + \lambda\|w\|_1, \quad \lambda = \frac{\beta}{2n\sigma_e^2}$$

Poll

A penalized negative log-likelihood has the form

$$-\ell(\boldsymbol{\theta}) + \text{pen}(\boldsymbol{\theta})$$

Example: logistic regression with quadratic penalty

True or false: Every estimate obtained by minimizing a penalized negative log-likelihood can be viewed as a MAP estimate. That is, for every penalty $\text{pen}(\boldsymbol{\theta})$ there is a prior $p(\boldsymbol{\theta})$ such that $\text{pen}(\boldsymbol{\theta}) = -\log p(\boldsymbol{\theta})$.

(A) True

(B) False

$\Leftrightarrow p(\theta) = e^{-\text{pen}(\theta)}$

$1 = \int p(\theta)\, d\theta = \int e^{-\text{pen}(\theta)}\, d\theta$

no reason to equal 1

$\theta > 1$

$\text{pen}(\theta) = \log\log\theta$

Multinomial Parameter Estimation

- Consider a pmf on K items given by $\theta_1, \ldots, \theta_K$
- After N iid draws, let

N_k = # of occurrences of outcome k

- Then we say $(N_1, \ldots, N_K)$ have a multinomial

  and write $(N_1, \ldots, N_K) \sim \text{multi}(N; \theta_1, \ldots, \theta_N)$

- MLE:

$$\hat{\theta}_k = \frac{N_k}{N}$$

Multinomial Parameter Estimation

- Dirichlet prior:

$$p(\theta_1, \ldots, \theta_K \mid \alpha_1, \ldots, \alpha_K) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

- Posterior:

$$p(\theta_1, \ldots, \theta_K \mid \alpha_1 + N_1, \ldots, \alpha_K + N_K)$$

$\alpha_k = 1 \; \forall k \Rightarrow$ uniform

- Posterior mode (MAP estimate):

$$\hat{\theta}_i = \frac{\alpha_i}{\sum \alpha_k}$$

- Posterior mean:

$$\hat{\theta}_i = \frac{\alpha_i - 1}{\sum \alpha_k - K}$$

$\alpha_i \rightarrow \alpha_i + N_i$

Final Thoughts

- When the posterior and prior distributions belong to the same family of distributions, the prior is said to be *conjugate* to the likelihood
- Priors are typically chosen to make posterior calculations easier
- Advantages of Bayesian estimation:
  - Natural confidence intervals
  - Make use of prior knowledge if available
- Disadvantages of Bayesian estimation:
  - Inference may be computationally demanding for complex models
  - Still requires tuning hyperparameters
  - May perform poorly if prior misspecified
- Next time: Gaussian processes for regression