Probabilistic Inference Basics

Gaurav Sharma

University of Rochester

Probabilistic Inference Problems

- We would like to make the "best" inference or "best" estimate of parameters
  - given observed data x and any prior knowledge we have
  - typically under assumed model
- What's best?
  - Inference: minimize probability of error E

$$\Pr(E) \tag{1}$$

- Estimation: most likely values of model parameters $\boldsymbol{\theta}$
  - Alternative formulations as minimization of an error metric

Optimal Decision

- Min $\Pr(E)$
- Possible decisions $1, 2, \ldots K$
- View from perspective of space of possible observations x
  - Need to partition space of possible observations into K decision regions $Z_1, Z_2, \ldots Z_K$
    - Decision i in region Z_i
  - Problem reduces to partitioning of possible space of observations
    - What areas should correspond to Z_i where decision is i?

MAP Decision Optimality

- Min $\Pr(E)$ equivalent to Max $\Pr(C)$ (Correct)

$$\begin{aligned}
\Pr(C) &= \sum_{i=1}^{K} \Pr(C, i) \\
&= \sum_{i=1}^{K} \Pr(C \mid i)\, p(i) \\
&= \sum_{i=1}^{K} \int_{Z_i} p(\mathrm{x} \mid i)\, p(i) d\mathrm{x}
\end{aligned}$$

- Optimal rule: For given observation x choose decision that maximizes the argument of integral

$$\hat{i} = \arg\max_{i} p(\mathrm{x} \mid i)\, p(i)$$

MAP Decision Optimality

- MAP Nomenclature
- Rule for Min $\Pr(E)$

$$\hat{i} = \arg\max_i p(\mathrm{x} \mid i)\, p(i) \equiv \arg\max_i p(i \mid \mathrm{x})$$

- A posteriori probability $p(i \mid \mathrm{x})$
- Maximum a posteriori probability (MAP) rule
  - MAP decisions minimize probability of error
- Intuitive: choose most likely decision given the data
- Computationally often challenging: likelihood and prior, prior often unknown
  - Sometimes likelihood is also challenging to formulate
- Why likelihood and prior partitioning?

MAP Estimation

- Model for data implicit in MAP decision
  - Including model parameters $\boldsymbol{\theta}$
- Often parameters are unknown a priori, need to also be estimated
- MAP estimates of parameters

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p\left(\boldsymbol{\theta} \mid \mathrm{x}\right) \equiv \arg\max_{\boldsymbol{\theta}} p\left(\mathrm{x} \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right)$$

- Intuitive: choose most likely value of parameters given the data
- Computationally often challenging: likelihood and prior, prior often unknown

Maximum Likelihood (ML) Decision and Parameter Estimation

- Prior often unknown in decision and estimation problems
  - Common assumption: equiprobable prior
- Recall: Posterior probability = likelihood $\times$ prior
- Under equiprobable prior: maximizing posterior probability $\equiv$ maximizing likelihood
  - Maximum likelihood (ML) decision/parameter estimation

MAP vs ML Decision

- MAP Decision

$$\hat{i} = \arg\max_i p\left(\mathrm{x} \mid i\right) p(i) \equiv \arg\max_i p\left(i \mid \mathrm{x}\right)$$

- ML Decision

$$\hat{i} = \arg\max_i p\left(\mathrm{x} \mid i\right)$$

- Likelihood often available from "forward" model
- MAP and ML decisions coincide for equiprobable priors

MAP vs ML Parameter Estimation

- MAP Estimate

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p\left(\boldsymbol{\theta} \mid \mathrm{x}\right) \equiv \arg\max_{\boldsymbol{\theta}} p\left(\mathrm{x} \mid \boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right)$$

- ML Estimate

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p\left(\mathrm{x} \mid \boldsymbol{\theta}\right)$$

- Likelihood often available from "forward" model
- MAP and ML estimates coincide for equiprobable priors

ML Parameter Estimation: Bernoulli Example

- X_i iid Bernoulli, with unknown parameter $\theta = \Pr\{X_i = 1\}$
- Observations $\mathrm{x} = [x_1, x_2, \ldots x_N]$ string of 0/1 values, length N
- ML Estimate of θ?
- Likelihood function $p\,(\mathrm{x} \mid \boldsymbol{\theta})$?
  - Example: $\mathrm{x} = 1101001001$, what is $p\,(\mathrm{x} \mid \boldsymbol{\theta})$?

ML Parameter Estimation: Bernoulli Example

- Likelihood function

$$p(\mathrm{x} \mid \boldsymbol{\theta}) = \theta^{\sum_i x_i}(1-\theta)^{N-\sum_i x_i} = \theta^{t(\mathrm{x})}(1-\theta)^{N-t(\mathrm{x})}$$

- Note $t(\mathrm{x}) = \sum_i x_i$ is a sufficient statistic
  - θ is conditionally independent of x given $t(\mathrm{x})$
- ML Estimate
  - Calculus of variations

$$t(\mathrm{x})\theta^{t(\mathrm{x})-1}(1-\theta)^{N-t(\mathrm{x})} - (N - t(\mathrm{x}))\theta^{t(\mathrm{x})}(1-\theta)^{N-t(\mathrm{x})-1} = 0$$

$$t(\mathrm{x})(1-\theta) - (N - t(\mathrm{x}))\theta = 0$$

$$\hat{\theta} = \frac{t(\mathrm{x})}{N}$$

- Intuitively appealing: probability of heads is estimated as the empirical fraction of observed heads

ML Parameter Estimation: Scalar Gaussian Example

- X_i iid Gaussian, with unknown mean μ and variance σ^2
- Parameters $\boldsymbol{\theta} = [\mu, \sigma^2]$
- Observations $\mathrm{x} = [x_1, x_2, \ldots x_N]$, sequence of real values
- ML Estimate of $\boldsymbol{\theta}$?
- Likelihood function $p(\mathrm{x} \mid \boldsymbol{\theta})$?

$$p(\mathrm{x} \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} p(x_i \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$\ln p(\mathrm{x} \mid \boldsymbol{\theta}) = \sum_{i=1}^{N} \left(-\frac{1}{2}\ln\left(2\pi\sigma^2\right) - \frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$= -\frac{N}{2}\ln\left(2\pi\sigma^2\right) - \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{2\sigma^2}$$

ML Parameter Estimation: Scalar Gaussian Example

- Log Likelihood function

$$\ln p\left(\mathrm{x} \mid \boldsymbol{\theta}\right) = -\frac{N}{2}\ln\left(2\pi\sigma^2\right) - \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{2\sigma^2}$$

- ML Estimates: $\hat{\boldsymbol{\theta}} = [\hat{\mu}, \hat{\sigma^2}] = \arg\max_{\boldsymbol{\theta}} \ln p\left(\mathrm{x} \mid \boldsymbol{\theta}\right)$
  - Calculus of variations

$$\hat{\mu} = \frac{1}{N}x_i$$

$$\hat{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \hat{\mu}\right)^2$$

- Intuitively appealing ML estimates are sample mean and sample variance

ML Parameter Estimation: Multivariate Gaussian

- Multivariate Gaussian X_i, d -dimensional iid Gaussian random vectors
  - with unknown parameters $\boldsymbol{\theta} = [\mu, \Sigma]$, $d \times 1$ mean vector μ and $d \times d$ covariance matrix Σ
- Observations $\mathrm{x} = [x_1, x_2, \ldots x_N]$, sequence of N, $d \times 1$ vectors
- Likelihood function

$$
\begin{aligned}
p\left(\mathrm{x} \mid \boldsymbol{\theta}\right) &= \prod_{i=1}^{N} \mathcal{N}(x_i | \mu, \Sigma) \\
&= \prod_{i=1}^{N} \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(-\frac{(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)}{2}\right) \\
\ln p\left(\mathrm{x} \mid \boldsymbol{\theta}\right) &= \sum_{i=1}^{N} \ln \mathcal{N}(x_i | \mu, \Sigma)
\end{aligned}
$$

ML Parameter Estimation: Multivariate Gaussian

- Log Likelihood function

$$
\begin{aligned}
\ln p\left(\mathrm{x} \mid \boldsymbol{\theta}\right) &= \sum_{i=1}^{N} \ln \mathcal{N}(x_i|\mu, \Sigma) \\
&= \sum_{i=1}^{N} \ln \left(\mathcal{N}(x_i|\mu, \Sigma)\right) \\
&= -\frac{N}{2} \ln \left((2\pi)^d \det(\Sigma)\right) \\
&\quad - \frac{1}{2} \sum_{i=1}^{N} \left((x_i - \mu)^T \Sigma^{-1} (x_i - \mu)\right)
\end{aligned}
$$

ML Parameter Estimation: Multivariate Gaussian

- Maximization of log likelihood
  - Tedious but straightforward
  - Matrix derivatives notation helps see Matrix Cookbook online
- ML Estimates
  - (Joint) solution of optimality conditions

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})(x_i - \hat{\mu})^T$$

- Intuitively appealing ML estimates are sample mean (vector) and sample covariance (matrix)

ML Parameter Estimation: Homogeneous Markov Chain

- X_i Homogeneous Markov Chain, with unknown parameters $\mathrm{p}^1 = [\Pr\{X_1 = i\}], P_{ij} = \Pr\{X_{n+1} = j \mid X_n = i\}$
- Observations $\mathrm{x} = [x_1, x_2, \dots x_N]$ string of values $\in \{1, 2, \dots L\}$, length N
- ML Estimate of θ?
- Likelihood function $p\left(\mathrm{x} \mid \boldsymbol{\theta}\right)$?

ML Parameter Estimation: Homogeneous Markov Chain: Example $L = 2$

- X_i Homogeneous Markov Chain, with unknown parameters $\mathrm{p}^1 = [\Pr\{X_1 = i\}], P_{ij} = \Pr\{X_{n+1} = j \mid X_n = i\}$
  - P is defined by α, β
  - Depending on assumptions also p^1
- Observations $\mathrm{x} = [x_1, x_2, \dots x_N]$ string of 0/1 values, length N

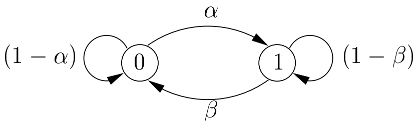

Figure: Transition Diagram representation of a Binary Markov Chain.

- ML Estimate of θ?
- Likelihood function $p\left(\mathrm{x} \mid \boldsymbol{\theta}\right)$?
  - Example: $\mathrm{x} = 1101001001$, what is $p\left(\mathrm{x} \mid \boldsymbol{\theta}\right)$?

ML Parameter Estimation: Homogeneous Markov Chain: Example $L = 2$

- Likelihood function $p(\mathrm{x} \mid \boldsymbol{\theta})$?
  - Example: $\mathrm{x} = 1101001001$, what is $p(\mathrm{x} \mid \boldsymbol{\theta})$?

$$p(\mathrm{x} \mid \boldsymbol{\theta}) = p^1(1)(1-\beta)\beta\alpha\beta(1-\alpha)\alpha\beta(1-\alpha)\alpha \tag{2}$$

$$= p^1(1)\alpha^{T_{0\to1}(\mathrm{x})}(1-\alpha)^{(N_0(\mathrm{x})-T_{0\to1}(\mathrm{x}))}\times \beta^{T_{1\to0}(\mathrm{x})}(1-\beta)^{(N_1(\mathrm{x})-T_{1\to0}(\mathrm{x}))} \tag{3}$$

- ML Estimates for α and β: Analogous to Bernoulli case

$$\hat{\alpha} = \frac{T_{0\to1}(\mathrm{x})}{N_0(\mathrm{x})} \tag{4}$$

$$\hat{\beta} = \frac{T_{1\to0}(\mathrm{x})}{N_1(\mathrm{x})} \tag{5}$$

- What about estimate of p^1?

ML Parameter Estimation: General Homogeneous Markov Chain

- X_i Homogeneous Markov Chain, with unknown parameters $\mathrm{p}^1 = [\Pr\{X_1 = i\}], P_{ij} = \Pr\{X_{n+1} = j \mid X_n = i\}$
- Observations $\mathrm{x} = [x_1, x_2, \ldots x_N]$ string of values $\in \{1, 2, \ldots L\}$, length N
- ML Estimate of transition probability matrix P:
  - Calculus of variations (with constraints)

$$\hat{P}_{ij} = \frac{T_{i \to j}(\mathrm{x})}{N_i(\mathrm{x})} \tag{6}$$

$T_{i \to j}(\mathrm{x}) = \#$ of $i \to j$ transitions in $\mathbf{x}$
$N_i(\mathrm{x}) = \#$ of transitions in x originating from state i

Outlook

- Using models for probabilistic inference and estimation
- Expectation Maximization
  - Modeling interactions that are not directly visible
  - Hidden/latent variables
  - Two case studies
    - WAMI to Roadmap alignment + Flow cytometry cell clustering
    - IID latent variables
  - General EM formulation and Gaussian mixture model

Outlook II

- Hidden Markov Models (HMMs)
  - Build upon Markov models
  - Memory + hidden/latent variables
  - Two case studies
    - Sequence alignment + Error correction decoding for convolutional codes
  - Three standard problems
    - Sequence estimation, individual state marginal probability estimation, parameter estimation
    - Parameter estimation: Baum-Welch $\equiv$ EM

Outlook III

- Stochastic context free grammars
  - Generalization of dependency beyond HMM Markovian structure
  - Example: Palindromes. Are they modeled well by a HMM?
  - Three standard problems analogous to HMMs
    - Sequence estimation (CYK), individual state marginal probability estimation, parameter estimation
    - Parameter estimation: Inside-Out $\equiv$ EM

Outlook IV

- Markov Random Fields
  - Generalization of dependency more appropriate for multi-dimensional data
  - Graphical representation
- Additional relevant examples from current ongoing research
  - Turbo decoding in communications and RNA structure prediction
  - Color barcodes