Density Estimation

ECE57000: Artificial Intelligence, Fall 2019

David I. Inouye

Announcements

- Resubmit HW2 only if you had formatting mistakes

<u>Maximum likelihood estimation (MLE)</u> is another way to estimate distribution parameters from samples

- **<u>Likelihood function</u>** how likely (or probable) a dataset $\mathcal{D} = \{x_i\}_{i=1}^n$ is under a distribution with parameters θ
$$\mathcal{L}(\theta; \mathcal{D}) = p(x_1, x_2, \dots, x_n; \theta)$$
- If we ***assume*** samples (or observations) of dataset are **<u>independent and identically distributed (iid)</u>**, then
$$\mathcal{L}(\theta; \mathcal{D}) = \prod_{i=1}^{n} p(x_i; \theta)$$
- Often simplified to the **<u>log-likelihood function</u>**
$$\ell(\theta; \mathcal{D}) = \log \mathcal{L}(\theta; \mathcal{D})$$
- Example: Coin flips with Bernoulli
- Non iid example: First flip Bernoulli, then alternating
- Example: Flight delays with exponential distribution

The <u>**likelihood function**</u> is a function of **parameters** θ as opposed to a density which is a function of x

- Sometimes written $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{x}) = p(\boldsymbol{x}; \boldsymbol{\theta})$

- Subtle but important difference with PDF/PMF
  - PDF/PMF are functions of x where θ is fixed
  - Likelihood is function of θ where x is fixed

- Additionally, likelihood function $\mathcal{L}$ is usually product of density functions (if **iid**)

<u>Maximum likelihood (MLE)</u> is another way to estimate distribution parameters from samples

- Optimize the following
$$\theta^* = \arg\max_\theta \mathcal{L}(\theta; \mathcal{D})$$
- (Derive negative log likelihood)
- Equivalent to
$$\theta^* = \arg\min_\theta -\frac{1}{n}\sum_{i=1}^{n} \log p(x_i; \theta)$$
- Wait, doesn't that look familiar?
- <u>**MLE equivalent to minimum KL divergence!**</u>

Example: Estimate Bernoulli parameter p given many coin flips

- $\mathcal{D} = \{H, T, T, T, H\}$

$$\theta^* = \arg\min_{\theta} -\frac{1}{n}\sum_{i=1}^{n} \log p(x_i; \theta)$$

Example: Estimate mean parameter λ of exponential distribution

- $\mathcal{D} = \{x_1, x_2, \dots, x_n\}$
- $p(x; \lambda) = \lambda \exp\{-\lambda x\}$
- $\log p(x; \lambda) = -\lambda x + \log \lambda$

$$\theta^* = \arg\min_{\theta} -\frac{1}{n} \sum_{i=1}^{n} \log p(x_i; \theta)$$

MLE is not always appropriate and fails in certain important situations

- Corrupt/noisy samples (related to **robustness**)
  - Cashiers using 1111 for birth year: 908 years old
  - One star ratings
- Finite (sometimes small) number of samples
  - One or two coin flips, Bernoulli
  - 1D with one sample, Gaussian
  - 2D with two samples, multivariate Gaussian

Robust estimators of a Gaussian mean can be computed using median

- Suppose corruption is 30% (e.g., 30% of cashiers don't put in correct birth year)
- MLE estimator of Gaussian is sample average

$$\arg\min_{\mu} \frac{1}{n} \sum_{i} \frac{1}{2} (x_i - \mu)^2$$

- Rather we can use the median which is:

$$\arg\min_{\mu} \frac{1}{n} \sum_{i} |x_i - \mu|$$

- (demo)

Regularized MLE is a way to handle finite or small sample sizes

- Maximize likelihood + regularization penalty
$$\arg\max_{\theta} \ell(\theta;\mathcal{D}) - \lambda R(\theta)$$
- Often written as minimizing negative likelihood
$$\arg\min_{\theta} -\ell(\theta;\mathcal{D}) + \lambda R(\theta)$$
- Concrete example for Gaussian mean estimation where $\sigma^2 = 1$ and $\lambda = \frac{1}{2}$
$$\arg\min_{\mu} -\ell(\mu;\mathcal{D}) + \frac{1}{2}\|\mu\|_2^2$$
$$\arg\min_{\mu} \sum_i \frac{1}{2}(x-\mu)^2 + \frac{1}{2}\mu^2$$

The most ubiquitous multivariate distribution is the multivariate Gaussian distribution

- Compare univariate to multivariate:
  - μ is mean and Σ is covariance

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\}$$

$$p(x_1, \dots, x_d) = \frac{1}{\left(\sqrt{2\pi}\right)^d \sqrt{\det \Sigma}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

- $\Theta = \Sigma^{-1}$ is called the **precision matrix** (or **inverse covariance**)
- Σ (and Θ) must be positive definite $\Sigma > 0$
- (Suppose $\Sigma = I$, suppose $\mu = 0$)

<u>Marginal</u> and <u>conditional</u> distributions are **Gaussian** and can be computed in **closed-form**

- 2D case:

$$\boldsymbol{x} = [x_1, x_2] \sim \mathcal{N}\left(\mu = [\mu_1, \mu_2], \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}\right)$$

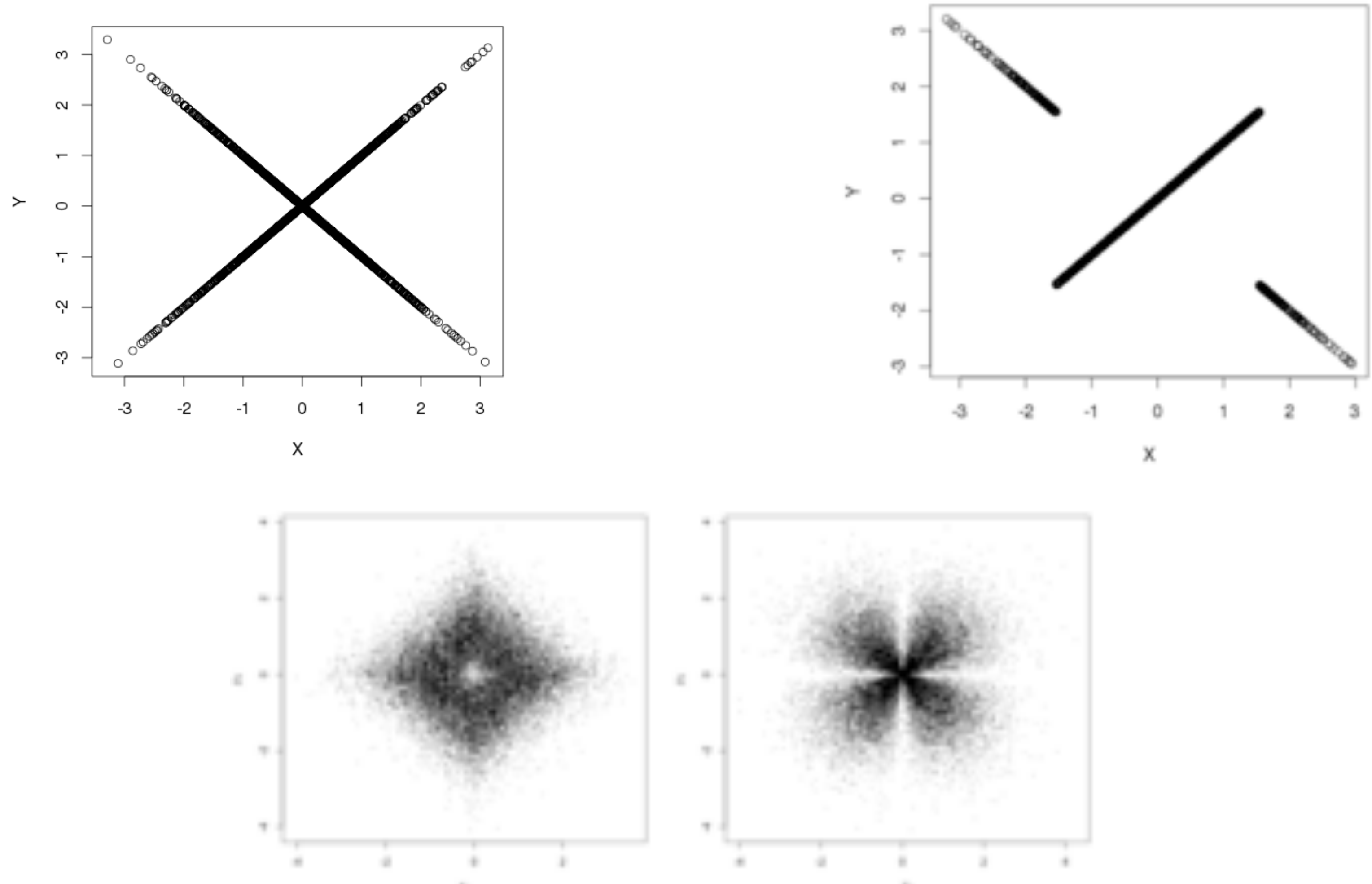
- Marginal distributions:

$$x_1 \sim \mathcal{N}(\mu = \mu_1, \sigma^2 = \sigma_1^2)$$
$$x_2 \sim \mathcal{N}(\mu = \mu_2, \sigma^2 = \sigma_2^2)$$

- Conditional distributions:

$$x_1 | x_2 = a$$
$$\sim \mathcal{N}\left(\mu = \mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(\mathrm{a} - \mu_2), \sigma^2 = \sigma_1^2 - \frac{\sigma_{21}^2}{\sigma_2^2}\right)$$

Gaussian marginals does NOT imply jointly multivariate Gaussian (converse NOT generally true)



MLE of multivariate Gaussian can be computed via empirical mean and covariance matrix

- Log-likelihood of multivariate Gaussian ($\mu = 0$)

$$-\frac{1}{2}\log|\Sigma| - \frac{1}{2n}\sum_{i=1}^{n} x_i^T \Sigma^{-1} x_i + const$$

- Three main identities:
  - $\frac{\partial \log|\mathrm{A}|}{\partial \mathrm{A}} = A^{-T}$
  - $\mathrm{Tr}(x^T A x) = \mathrm{Tr}(A x x^T)$
  - $\frac{\partial \mathrm{Tr}(AX)}{\partial \mathrm{X}} = A$
- Hint: Do derivative with respect to Σ^{-1}