Density Estimation

ECE57000: Artificial Intelligence, Fall 2019
David I. Inouye
Announcements

- Quiz 4 on Wednesday
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

\[
\text{arg min}_\mu \frac{1}{n} \sum_i \frac{1}{2} (x_i - \mu)^2
\]

- Rather we can use the median which is:

\[
\text{arg min}_\mu \frac{1}{n} \sum_i |x_i - \mu|
\]

- (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 \( \lambda = \frac{1}{2} \)

  \[ \arg \min_{\mu} -\ell(\mu; \mathcal{D}) + \frac{1}{2} \| \mu \|_2^2 \]

  \[ \arg \min_{\mu} \sum_{i} \frac{1}{2} (x_i - \mu)^2 + \frac{1}{2} \mu^2 \]
Derivation for regularized Gaussian mean estimation

\[
L \left( \mu; \mathcal{D}, \lambda = \frac{1}{2} \right) = -\ell(\mu; \mathcal{D}) + R(\mu) = \sum_i \frac{1}{2} (x_i - \mu)^2 + \frac{1}{2} \mu^2
\]

\[
\frac{\partial L}{\partial \mu} = \sum_i \frac{1}{2} \left( 2(x_i - \mu) \right)(-1) + \frac{1}{2} (2\mu)
\]

\[
= \mu + \sum_i (\mu - x_i) = \mu + n\mu - \sum_i x_i
\]

\[
\frac{\partial L}{\partial \mu} = 0 = (1 + n)\mu - \sum_i x_i
\]

\[
\mu = \frac{1}{n+1} \sum_i x_i
\]
The most ubiquitous multivariate distribution is the **multivariate Gaussian/normal distribution**

- Compare univariate to multivariate:
  - $\mu$ is mean and $\Sigma$ is covariance

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

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

- $\Theta = \Sigma^{-1}$ is called the **precision matrix** (or **inverse covariance**)
- $\Sigma$ (and $\Theta$) must be positive definite $\Sigma > 0$
- (Suppose $\Sigma = I$, suppose $\mu = 0$)
Multivariate Gaussian is independent “spherical” Gaussian that is rotated and scaled

\[
\Sigma = U \Lambda U^T = (U \Lambda^{\frac{1}{2}})(\Lambda^{\frac{1}{2}} U^T) = (U \Lambda^{\frac{1}{2}})(U \Lambda^{\frac{1}{2}})^T
\]

\[
x^T (U \Lambda^{-\frac{1}{2}})(U \Lambda^{-\frac{1}{2}})^T x = (\Lambda^{-\frac{1}{2}} U x)^T (\Lambda^{-\frac{1}{2}} U x) = z^T z
\]

**Figure 4.1** Visualization of a 2 dimensional Gaussian density. The major and minor axes of the ellipse are defined by the first two eigenvectors of the covariance matrix, namely \(u_1\) and \(u_2\). Based on Figure 2.7 of (Bishop 2006a).
Marginal and conditional distributions are Gaussian and can be computed in closed-form.

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

- **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:**
  \[ x_1 | x_2 = a \sim \mathcal{N} \left( \mu = \mu_1 + \frac{\sigma_{12}}{\sigma_2^2} (a - \mu_2), \sigma^2 = \sigma_1^2 - \frac{\sigma_{21}^2}{\sigma_2^2} \right) \]
Gaussian marginals does **NOT** imply jointly multivariate Gaussian (converse **NOT** generally true)
Affine transformations of multivariate Gaussian vector are also multivariate Gaussian

- If $x \sim \mathcal{N}(\mu, \Sigma)$ and $y = Ax + b$, then $y \sim \mathcal{N}(A\mu + b, A\Sigma A^T)$.

- Special case: Marginal distribution when $A$ is:

$$A_i = \begin{cases} 
1, & \text{if } i = k \\
0, & \text{otherwise}
\end{cases}$$

then $y = x_k \sim p(x_k)$.

- Key point: Marginals, conditionals and affine functions known in **closed-form**.

- Consequence 1: Easy to manipulate.

- Consequence 2: Gaussians and linear ideas play nicely with each other.
MLE of multivariate Gaussian can be computed via empirical mean and covariance matrix

- Log-likelihood of multivariate Gaussian ($\mu = 0$)
  \[
  \mathcal{L}(\Sigma; D) = \sum_{i=1}^{\mathcal{D}} \left[ -\frac{1}{2} x_i^T \Sigma^{-1} x_i - \frac{1}{2} \log|\Sigma| + \frac{d}{2} \log 2\pi \right]
  \]

- Three main identities:
  1. \[ \frac{\partial \log|A|}{\partial A} = A^{-T} \]
  2. \[ \text{Tr}(x^T A x) = \text{Tr}(A x x^T) \]
  3. \[ \frac{\partial \text{Tr}(A X)}{\partial X} = A \]
  4. Hint: Do derivative with respect to $\Sigma^{-1}$
Simplification and derivation of MLE for multivariate Gaussian

\[ L(\Sigma; \mathcal{D}) = \frac{n}{2} \log |\Sigma^{-1}| - \frac{1}{2} \text{Tr} \left( \Sigma^{-1} \left( \sum_i x_i x_i^T \right) \right) \]

\[ \frac{\partial L}{\partial \Sigma^{-1}} = \frac{n}{2} \Sigma - \frac{1}{2} \sum_i x_i x_i^T \]

\[ \Sigma = \frac{1}{n} \sum_i x_i x_i^T \]