Lecture 9 - Fisher LDA + Gaussian MLE

Today: Fisher LDA
- math tricks & MLE for Gaussian

- generative model for classification: (Fisher) linear discriminant analysis
  \[ ELD \text{ (instead of LDA)} \]

- for classification \( y \in \{0, 1\} \)
  \[ x \in \mathbb{R}^d \text{ (class condition)} \]

- generative approach: \( p(x, y; \theta) = p(x | y; \theta) p(y; \phi) \)

- vs.

- conditional approach: \( p(y | x; \theta) \)

- For Fisher model: we assume \( p(x | y; \theta) = \mathcal{N}(x | \mu_y, \Sigma) \)

- \[ \theta = (\mu_0, \mu_1, \Sigma, \pi) \]
  - mean of class 0
  - showed \( p(y=1) \)

- as before (see exponential family argument)

- can show that \( p(y | x; \theta) = \mathcal{G}(w^T x) \) where \( \mathcal{G} \) is a \( \mathcal{G}(\mu, \Sigma, \phi, \theta) \)

- [note: if use \( \phi = \mathcal{G}(\mu, \Sigma) \), get "quadratic discriminant analysis"

- i.e. \( \mathcal{G}(w^T y, \Sigma) \) where \( \mathcal{G}(\mu, \Sigma) \) is a quadratic function

- \[ \mathcal{G}(w^T y, \Sigma) \]
  - \[ \Sigma = \mathcal{G}(w^T y, \Sigma) \]
  - \[ \Sigma = \mathcal{G}(w^T y, \Sigma) \]

- gen. approach: do joint MLE to estimate

- \[ \mathcal{G}(w^T y, \Sigma) \]

- [vs. \( \min \mathcal{G}(w^T y, \Sigma) \) for logistic regression]

- \[ x \sim \mathcal{N}(\mu, \Sigma) \]
  - \[ \mu \in \mathbb{R}^d \]
  - \[ \Sigma \in \mathbb{R}^{d \times d} \]

- \[ \mathcal{G}(w^T y, \Sigma) \]
$\mu \in \mathbb{R}^d$ and
$\Sigma \in \mathbb{R}^{d \times d}$ is symmetric
$\Sigma > 0$

$E[v] = \frac{1}{\Sigma} \left( v^T (x - \mu) \right) \frac{1}{(x - \mu)^T} \Sigma
$\right) > 0$

$h(A) = h(B)$

$\Delta = (\mu, \Sigma)$

$\Delta \Sigma^{-1} \Delta^T > 0$

$log \text{-} \text{likelihood} = \sum_{i=1}^{n} \log p(x_i) = \text{const} - \frac{1}{2} \log |\Sigma| - \frac{1}{2} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)$

$\Delta \Sigma^{-1} \frac{1}{2} \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \log |\Sigma|$

vector derivative review:

Suppose $f: \mathbb{R}^m \to \mathbb{R}^n$

If $f$ is $C^1$ at $x_0$ then $\|Df(x_0)\|_2 = o(\|x - x_0\|)$

$s$ is differentiable at $x_0$ if $s$ is a linear operator $ds_x: \mathbb{R}^n \to \mathbb{R}^m$

$s(x + d) - s(x) = ds_x(d) + o(\|d\|)$

1) $ds_x(\Delta) + b ds_x(\Delta) = ds_x(\Delta + b)$

2) be careful with dimensions

$\Delta: \mathbb{R}^m \to \mathbb{R}^n$

$ds_x$ is a row vector, $(1 \times m)$
\[ \frac{\partial (g \circ f)_{x_0}}{\partial x} = \frac{\partial g_{y_0}}{\partial y} \cdot \frac{\partial f_{x_0}}{\partial x} \]

\[ g(f(x_0)) = (g(f(x_0))) \text{ matrix product of Jacobians} \]

\[ f(x) = x - \mu \]
\[ \frac{\partial f_{x_0}}{\partial x} = -I \]
\[ g(w) = w^T A w \]
\[ \frac{\partial g_{w_0}}{\partial w} = w^T (A + A^T) \]
\[ g - f (\mu) = (x - \mu)^T A (x - \mu) \]
\[ \frac{\partial (g - f)_{\mu_0}}{\partial \mu} = \frac{\partial g_{w_0}}{\partial w} \cdot \frac{\partial f_{x_0}}{\partial x} = (x - \mu)^T (A + A^T) (-I) \]

for Gaussian:
\[ \nabla A : \frac{-1}{2} E(i = 1 \leq x_i - \mu) \]
\[ \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \]

\[ \hat{\Sigma} \]
\[ m_{MLE} = \frac{1}{\begin{bmatrix} \Sigma \end{bmatrix}} \]

\[ \left( \begin{array}{c} x_1 \\ \vdots \\ x_n \end{array} \right) \]

15h 58

Example 2: derivative of \( f(A) = \log \det(A) \) where assume \( A \) is symmetric
\[ A > 0 \]

\[ \log \det (A + \Delta) - \log \det (A) \]
\[ = \log \det (A^{1/2} (I + A^{-1/2} \Delta A^{-1/2}) A^{1/2}) - \log \det (A) \]
\[ = \log |A^{1/2} (I + A^{-1/2} \Delta A^{-1/2}) A^{1/2}| - \log |A| \]
\[ = \log \left| \sum \lambda_i \right| \text{ e-value of } \Delta \]
\[ = \sum \log \lambda_i (I + A^{-1/2} \Delta A^{-1/2}) \]
\[ \lambda (I + B) = 1 + \lambda (B) \]
\[ \lambda (I + B) = B \]

\[ \log (1 + \lambda^2) = \lambda + o(\lambda^2) \text{ for } \lambda \ll 1 \]

\[ \log \left| \sum \lambda_i \left( A^{-1/2} \Delta A^{-1/2} \right) \right| \]
\[ = \sum \log \left( A^{-1/2} \Delta A^{-1/2} \right) \]

\[ = \log \left( A^{-1/2} \Delta A^{-1/2} \right) + o(1) \]

\[ A \text{ is homogeneous set.} \]

\[ \left( \begin{array}{c} \mathbf{B} \\ \mathbf{v} \end{array} \right) \]

\[ \mathbf{B} = A \]

\[ \mathbf{v} = \dot{\mathbf{a}} \]

\[ \mathbf{v} = \dot{\mathbf{a}} \]
see Boyd’s book A.4.1 for the above proof

\[ \log \det (A) = A^{-1} \]

Back to log-Affine of Gaussian:

\[ + \frac{n}{2} \log |\Sigma - \frac{1}{n} \mathbb{E}(\Sigma) > \quad \text{(concave, opt. of } -\lambda = \mathbb{E}^{-1}) \]

Take derivatives w.r.t.

\[ \mathbb{E}^{-1} = \Lambda \text{ s.t. } \frac{n}{2} (\mathbb{E}^{-1})^T - \frac{n}{2} \mathbb{E} (\mu) = 0 \]

\[ \Rightarrow \mathbb{E}_{\text{MLE}} = \mathbb{E} (\mu_{\text{MLE}}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_{\text{MLE}})(x_i - \mu_{\text{MLE}})^T \]

(The empirical covariance matrix)

Unsupervised learning

Here \( X \) without any label \( Y \)

Consider the Gaussian mixture model (GMM)

\( Y \sim \text{mult}(\pi) \quad \pi \in \Delta K \)

\( X | Y = k \sim N(\mu_k, \Sigma) \)

\( p(x) = \sum_{y} p(x, y) = \sum_{y} p(x | y)p(y) = \sum_{y} \pi_y N(x | \mu_y, \Sigma) \)

\( \text{“GMM model”} \quad \text{(more generally, can have } \Sigma \text{ per class)} \)

Graphical model for this “latent variable model”
two values on $p(x)$ <
shaclen
\[ p(x) = \frac{1}{Z_0} p_0(x(z_1)) p_0(x(z_0)) \]

later in class: we will call these Shaclen $H_{**}^{**}
\[
\begin{align*}
2x & \rightarrow x_1 & \rightarrow x_2 \\
& \rightarrow x_3 & \rightarrow x_4 \\
& \rightarrow x_5 & \rightarrow x_6
\end{align*}
\]

K-means → to do clustering i.e. Group data

we want to get a cluster assignment for every data pt. $X_i$

\[ z_i,j = 1 \text{ if } x_i \text{ belongs to cluster } j \]

\[ j = 1, \ldots, K \text{ # of clusters (specified in advance) for K-means} \]

Applications:

- Vector quantization (compression)
- In computer vision: use K-means to get "bag of visual words" representation of image patches
- Many many others