Foundations of Machine Learning
AI2000 and AI5000

FoML-13
Probabilistic Generative Models - Continuous features

Dr. Konda Reddy Mopuri
Department of AI, IIT Hyderabad
July-Nov 2025

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

DiL
Data-driven Intelligence
& Learning Lab

So far in FoML

- What is ML and the learning paradigms
- Probability refresher
- MLE, MAP, and fully Bayesian treatment
- Linear Regression with basis functions - regularization & model selection
- Bias-Variance Decomposition/Tradeoff (Bayesian Regression)
- Decision Theory - three broad classification strategies

Probabilistic Generative Models

Probabilistic Generative Models (K=2)

- Goal is to recover
  - Class conditional densities - $p(x/C_k)$
  - Prior densities - $p(C_k)$
  - → Joint distribution - $p(x, C_k) = p(x/C_k)\, p(C_k)$
  - → Posterior distribution

$$p(C_1|\mathbf{x}) = \frac{p(x/C_1)\, p(C_1)}{p(x)}$$

$p(x) \longrightarrow p(x/C_1)\, p(C_1) + p(x/C_2) \cdot p(C_2)$

Probabilistic Generative Models (K=2)

$$p(\mathcal{C}_1|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}$$

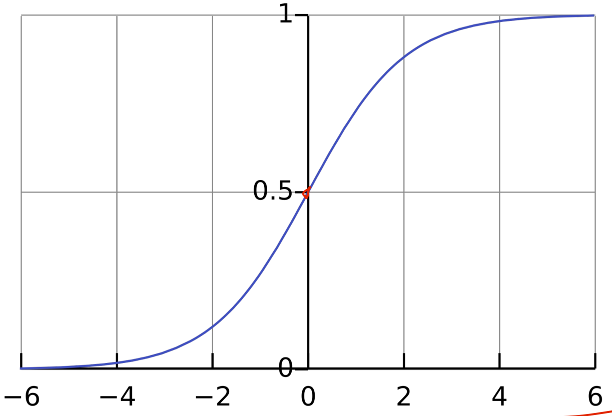
$$= \frac{1}{1+\exp(-a)} = \underbrace{\sigma(a)}_{\text{logistic sigmoid}(a)}$$

$$a = \ln \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}$$

Logit function (log odds)

Logistic Sigmoid

$\sigma(a) = P(C_1|x)$



$\to a$

as $a \Uparrow$ $\sigma(a) \to 1$
$p(C_2|x) \to 0$

$a \Downarrow$ $\sigma(a) \to 0$
$p(C_2|x) \to 1$

- S-shaped
- Squashing function

$$\sigma(-a) = 1 - \sigma(a)$$

$$\sigma'(a) = \sigma(a)[1-\sigma(a)]$$

Probabilistic Generative Models (K>2)

- For multiple classes

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(x/C_k)\ p(C_k)}{\sum_{j=1}^{K} p(x/C_j)\ p(C_j)}$$

$$a_k = \ln[p(x/C_k) \cdot p(C_k)]$$

$$p(C_k|x) = \frac{e^{a_k}}{\sum_{i=1}^{K} e^{a_i}}$$

Normalized exponential (multiclass generalization of sigmoid)

Also, known as 'softmax'

$$a_k \gg a_j \quad \forall j \neq k \qquad p(C_k|x) \rightarrow 1$$

Let's choose specific forms for the class conditional densities

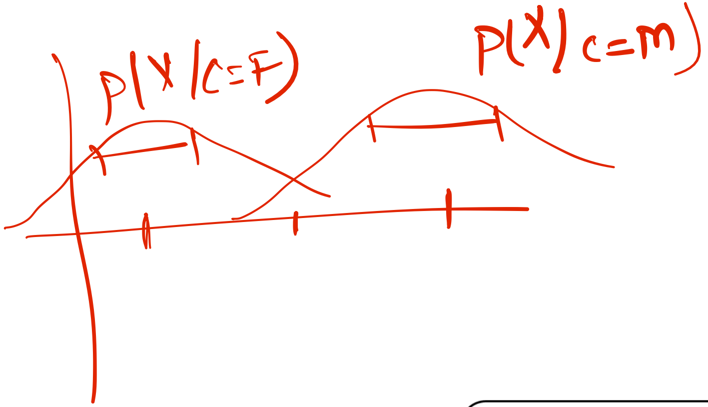
Class conditional densities: Continuous i/p

- Gaussian class conditional densities

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\}$$

- Assume shared covariance matrix

$\Sigma_k = \Sigma \quad \forall \; k=1,\ldots K$



Class conditional densities: Continuous i/p

- 2 classes case

$$p(\mathcal{C}_1/\mathbf{x}) = \frac{p(x/C_1)\,p(C_1)}{p(x/C_1)\,p(C_1) + p(x/C_2)\,p(C_2)}$$

$$= \frac{1}{1+e^{-a}} = \sigma(a)$$

$$a = \ln\left[\frac{p(x/C_1)\cdot p(C_1)}{p(x/C_2)\cdot p(C_2)}\right]$$

$$= \underbrace{\ln\left[\frac{p(x/C_1)}{p(x/C_2)}\right]} + \ln\left[\frac{p(C_1)}{p(C_2)}\right]$$

$$p(x/C_k) = N(x \mid \mu_k, \Sigma)$$

$$= \ln\left[e^{-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)^T \Sigma^{-1}(x-\mu_2)}\right] + \ln\left[\frac{p(C_1)}{p(C_2)}\right]$$

$$= -\frac{1}{2}\left[\cancel{x^T \Sigma^{-1} x} - 2x^T \Sigma^{-1}\mu_1 + \mu_1^T \Sigma^{-1}\mu_1 - \cancel{x^T \Sigma^{-1} x} + 2x^T \Sigma^{-1}\mu_2 - \mu_2^T \Sigma^{-1}\mu_2\right] + \ln\left[\frac{p(C_1)}{p(C_2)}\right]$$

$$a = (\mu_1 - \mu_2)^T \Sigma^{-1} x - \frac{1}{2}\mu_1^T \Sigma^{-1}\mu_1 + \frac{1}{2}\mu_2^T \Sigma^{-1}\mu_2 + \ln\left[\frac{p(C_1)}{p(C_2)}\right]$$

$$= w^T x + w_0$$

Class conditional densities: Continuous i/p

- 2 classes case
- Shared covariance → $\boxed{\text{Linear Discriminant}}$ and Generalized linear model

$p(C_1|x) = p(C_2|x)$

Decision boundary

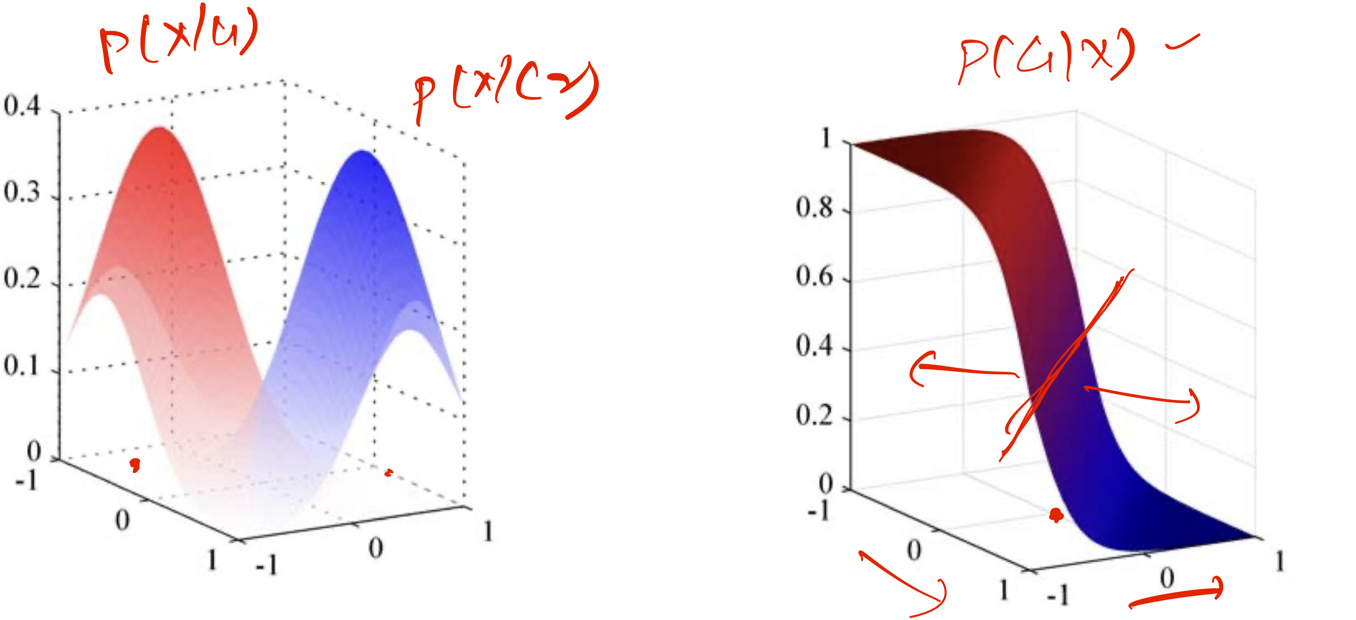
$a = \ln [\cdot]$

$= 0$

$\sigma(a) = 1/2$

$$\begin{aligned} \mathbf{w} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ w_0 &= -\frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}. \end{aligned}$$

Class conditional densities: Continuous i/p



Left: Gaussian class conditional densities Right: Posterior Probability for the Red class (logistic sigmoid of a linear function of i/p x)

Class conditional densities: Continuous i/p

- General case (K>2)

$$a_k(\mathbf{x}) = \mathbf{w}_k^{\mathrm{T}}\mathbf{x} + w_{k0}$$

$$\begin{aligned} \mathbf{w}_k &= \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k \\ w_{k0} &= -\frac{1}{2}\boldsymbol{\mu}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln p(\mathcal{C}_k) \end{aligned}$$

softmax $\quad \dfrac{e^{a_k}}{\sum_{j=1}^{K} e^{a_j}} = \dfrac{e^{a_k + \text{const}}}{\sum_{j=1}^{K} e^{a_k + \text{const}}}$

$$a_k(x) = \ln\left[p(x/C_k) \cdot p(C_k)\right]$$

$$= \ln p(x|C_k) + \ln\left[p(C_k)\right]$$

$$= \ln\left[\frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)}\right] + \ln\left[p(C_k)\right]$$

$$= \underbrace{\ln\left[\frac{1}{(2\pi)^{D/2}}\cdot\frac{1}{|\Sigma|^{1/2}}\right]}_{\text{Constant independent of } k\text{; same } \forall a_j} - \frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k) + \ln\left[p(C_k)\right]$$

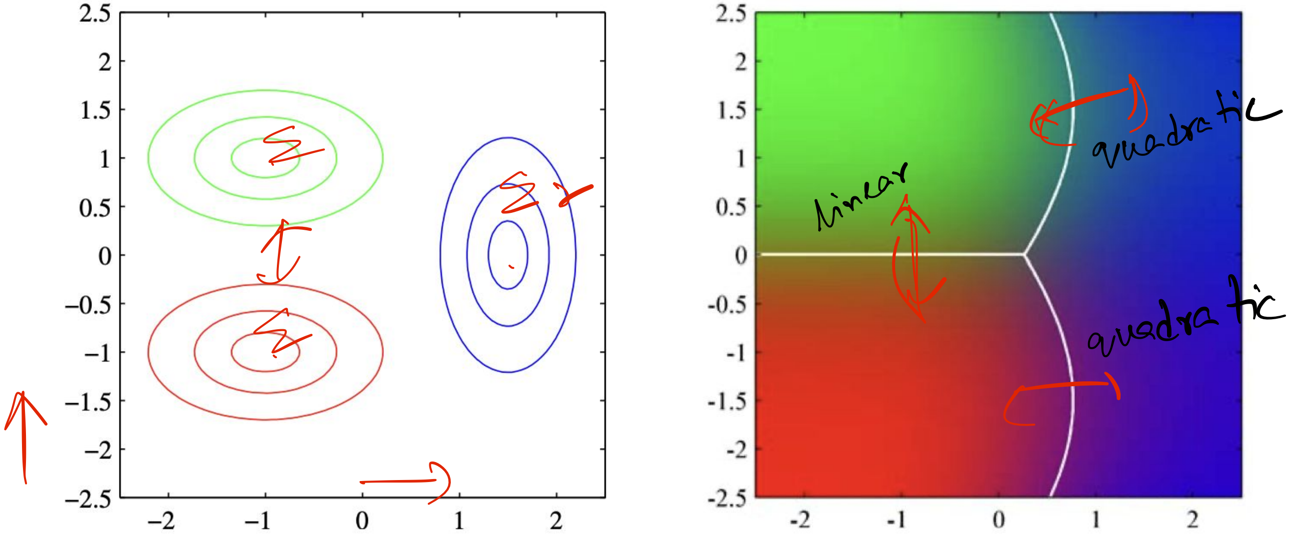
$$= -\frac{1}{2}\left[\underbrace{x^T\Sigma^{-1}x}_{\text{same } \forall a_j} - 2x^T\Sigma^{-1}\mu_k + \mu_k^T\Sigma^{-1}\mu_k\right] + \ln\left[p(C_k)\right]$$

$$= \underbrace{(\Sigma^{-1}\mu_k)^T}_{w_k} x \underbrace{- \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \ln\left[p(C_k)\right]}_{w_{k0}}$$

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Class conditional densities: Continuous i/p

General case (K>2)



Left: Gaussian class conditional densities (G and R have same covariance but B different) Right: Posterior Probabilities for the all the classes (corresponding RGB vector components)

Maximum Likelihood

LDA: MLE for K=2

- Dataset: input $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

  Binary targets $\mathbf{t} = \{t_1, \ldots, t_N\}$

$t_n = \{0, 1\}$

LDA: MLE for K=2

- Gaussian conditional densities $$p(\mathbf{x}|C_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\{\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\}$$
- Use MLE to estimate
  - μ_k, Σ, and priors $p(C_k)$
- Denote the priors with π and $1-\pi$

$p(X/C_k)$

C.C.D

$P(C_1) = \pi$

$P(C_2) = 1-\pi$

priors

For x_n with $t_n = 1$: $p(\mathbf{x}_n, C_1) =$ $p(x_n|C_1)\, p(C_1)$

For x_n with $t_n = 0$: $p(\mathbf{x}_n, C_2) =$ $p(x_n|C_2)\, p(C_2)$

LDA: MLE for K=2

The likelihood is given by (assuming iid data)

X

$$p(\mathbf{t}|\pi, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = \prod_{n=1}^{N} [\pi \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_1, \boldsymbol{\Sigma})]^{t_n} [(1-\pi)\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_2, \boldsymbol{\Sigma})]^{1-t_n}$$

$$= \prod_{n=1}^{N} P(x_n, t_n) = \prod_{n=1}^{N} P(x_n|t_n)\, P(t_n)$$

LDA: MLE for K=2

Consider the log likelihood

$$\ln p(\mathbf{t}, \mathbf{X}/\pi, \mu_1, \mu_2, \Sigma) = \sum_{n=1}^{N} t_n \ln \pi + t_n \ln \mathcal{N}(\mathbf{x}_n/\mu_1, \Sigma) + (1 - t_n) \ln (1 - \pi) + (1 - t_n) \ln \mathcal{N}(\mathbf{x}_n/\mu_2, \Sigma)$$

LDA: MLE for K=2

Estimate for π

$$\ln p(\mathbf{t}, \mathbf{X}/\pi, \mu_1, \mu_2, \Sigma) = \sum_{n=1}^{N} t_n \ln \pi + t_n \ln \mathcal{N}(\mathbf{x}_n/\mu_1, \Sigma) +$$
$$(1 - t_n) \ln (1 - \pi) + (1 - t_n) \ln \mathcal{N}(\mathbf{x}_n/\mu_2, \Sigma)$$

$$\frac{\partial}{\partial \mu} \left(\quad \right) = 0$$

LDA: MLE for K=2

Estimate for μ_1

$$\ln p(\mathbf{t}, \mathbf{X}/\pi, \mu_1, \mu_2, \Sigma) = \sum_{n=1}^{N} t_n \ln \pi + t_n \ln \mathcal{N}(\mathbf{x}_n/\mu_1, \Sigma) + (1 - t_n) \ln (1 - \pi) + (1 - t_n) \ln \mathcal{N}(\mathbf{x}_n/\mu_2, \Sigma)$$

$$\boldsymbol{\mu}_1 = \frac{1}{N_1} \sum_{n=1}^{N} t_n \mathbf{x}_n$$

$$\frac{\partial}{\partial \mu_1}\left(\cdot \right) = 0$$

LDA: MLE for K=2

Estimate for μ_2

$$\ln p(\mathbf{t}, \mathbf{X}/\pi, \mu_1, \mu_2, \Sigma) = \sum_{n=1}^{N} t_n \ln \pi + t_n \ln \mathcal{N}(\mathbf{x}_n/\mu_1, \Sigma) +$$
$$(1 - t_n) \ln (1 - \pi) + (1 - t_n) \ln \mathcal{N}(\mathbf{x}_n/\mu_2, \Sigma)$$

$$\boldsymbol{\mu}_2 = \frac{1}{N_2} \sum_{n=1}^{N} (1 - t_n) \mathbf{x}_n$$

LDA: MLE for K=2

Estimate for Σ

$$\ln p(\mathbf{t}, \mathbf{X}/\pi, \mu_1, \mu_2, \Sigma) = \sum_{n=1}^{N} t_n \ln \pi + t_n \ln \mathcal{N}(\mathbf{x}_n/\mu_1, \Sigma) + (1-t_n)\ln(1-\pi) + (1-t_n)\ln \mathcal{N}(\mathbf{x}_n/\mu_2, \Sigma)$$

$$\Sigma_{ML} = \frac{N_1}{N}\left[\frac{1}{N_1}\Sigma_{n=1}^{N} t_n(\mathbf{x}_n - \mu_{\mathbf{1,ML}})(\mathbf{x}_n - \mu_{\mathbf{1,ML}})^T\right] + \frac{N_2}{N}\left[\frac{1}{N_2}\Sigma_{n=1}^{N}(1-t_n)(\mathbf{x}_n - \mu_{\mathbf{2,ML}})(\mathbf{x}_n - \mu_{\mathbf{2,ML}})^T\right]$$

Weighted average of the sample covariances

LDA: MLE for K=2

The ML solutions

LDA: MLE for K=2

The posterior for a new data point x'

$$p(C_1/\mathbf{x}') = \sigma(\mathbf{w}_{ML}^T\mathbf{x}' + w_{0,ML})$$

$$\mathbf{w}_{ML} = \Sigma_{ML}^{-1}(\mu_{1,ML} - \mu_{2,ML})$$

$$w_{0,ML} = -\frac{1}{2}\mu_{1,ML}^T\Sigma_{ML}^{-1}\mu_{1,ML} + \frac{1}{2}\mu_{2,ML}^T\Sigma_{ML}^{-1}\mu_{2,ML} + \ln\frac{\pi_{ML}}{1-\pi_{ML}}$$

Next
PGM for discrete data
Discriminant Functions