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

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

Data-driven Intelligence & Learning Lab

Probabilistic Generative Models (K=2)

- Goal is to recover
  - Class conditional densities -
  - Prior densities -
  - → Joint distribution -
  - → Posterior distribution

$$p(\mathcal{C}_1|\mathbf{x}) \quad =$$

Probabilistic Generative Models (K=2)

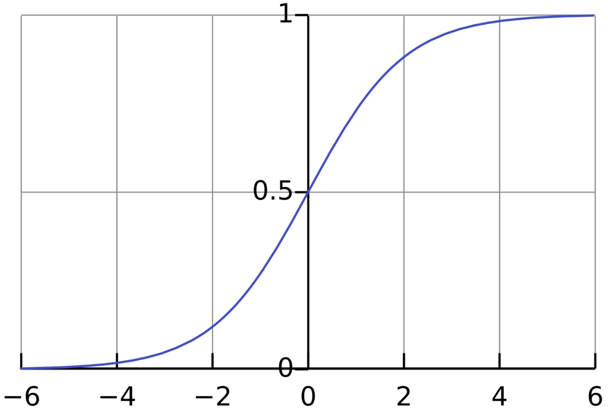
$$p(\mathcal{C}_1|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1) + p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}$$

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

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

Logit function (log odds)

Logistic Sigmoid



- S-shaped
- Squashing function

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

Probabilistic Generative Models (K>2)

- For multiple classes

$$p(\mathcal{C}_k|\mathbf{x}) \quad =$$

Normalized exponential (multiclass generalization of sigmoid)

Also, known as 'softmax'

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

Class conditional densities: Continuous i/p

- 2 classes case

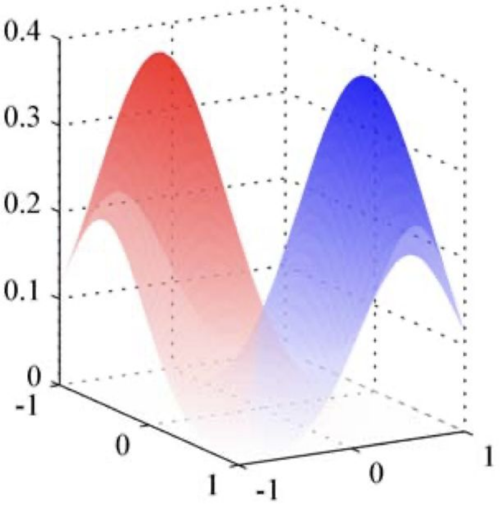
$$p(\mathcal{C}_1/\mathbf{x}) =$$

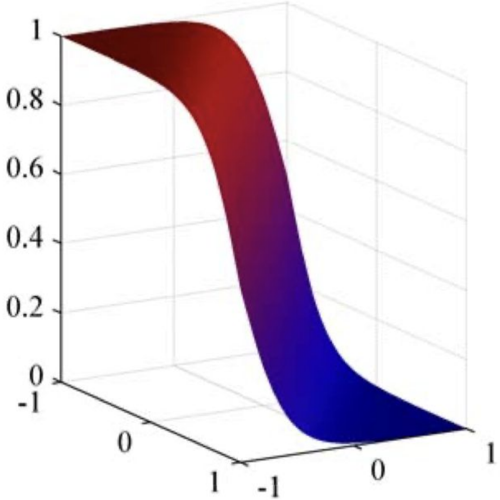
Class conditional densities: Continuous i/p

- 2 classes case
- Shared covariance → Linear Discriminant and Generalized linear model

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

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

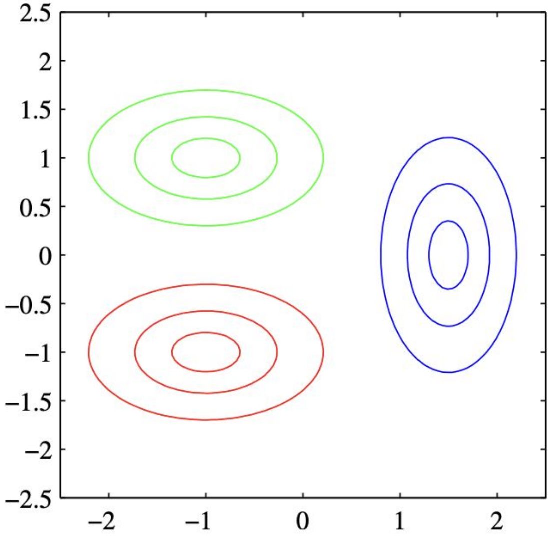
- General case (K>2)

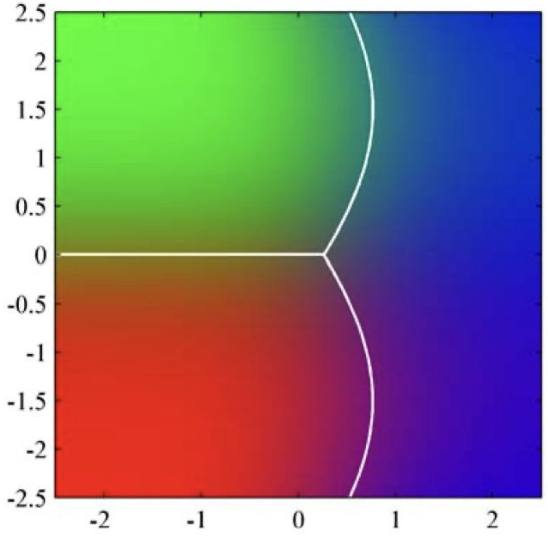
$$a_k(\mathbf{x}) = \mathbf{w}_k^{\mathrm{T}}\mathbf{x} + w_{k0}$$

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

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 $\mathbf{t} = \{t_1, \ldots, _N\}$

LDA: MLE for K=2

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

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

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

LDA: MLE for K=2

The likelihood is given by (assuming iid data)

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

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

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

LDA: MLE for K=2

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

Estimate for μ_1

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

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

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

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

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