Intro to ML

October 6th, 2021

Regression

measurement

Input x, f is what we want

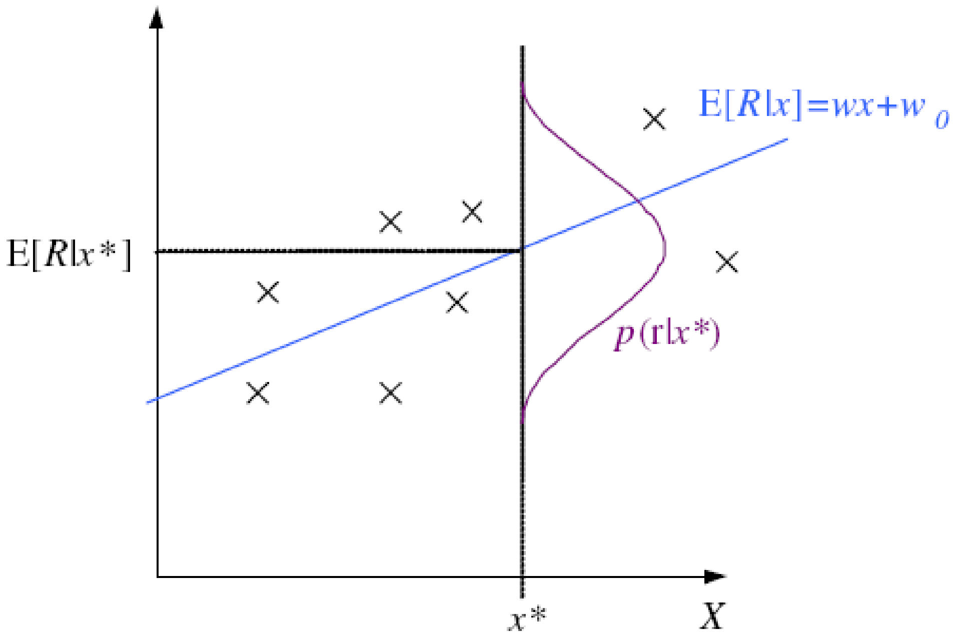
$$r = f(x) + \varepsilon$$

$$\text{estimator}: g(x \mid \theta)$$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

$$\boxed{p(r \mid x) \sim \mathcal{N}(g(x \mid \theta), \sigma^2)}$$

A value out of g() add to a Gaussian distribution -> a Gaussian distribution with mean coming from g(x)



$$\mathcal{L}(\theta \mid \mathcal{X}) = \log \prod_{t=1}^{N} p(x^t, r^t)$$

Joint = condition x unconditioned

$$= \log \prod_{t=1}^{N} p(r^t \mid x^t) + \cancel{\log \prod_{t=1}^{N} p(x^t)}$$

Likelihood function

Regression: From LogL to Error

Expected value

$$\mathcal{L}(\theta \mid \mathcal{X}) = \log \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{\left[r^t - \boxed{g\left(x^t \mid \theta\right)}\right]^2}{2\sigma^2}\right]$$

$$= -\cancel{N \log \sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{t=1}^{N} \left[r^t - g\left(x^t \mid \theta\right)\right]^2$$

$$E(\theta \mid \mathcal{X}) = \frac{1}{2} \sum_{t=1}^{N} \left[r^t - g\left(x^t \mid \theta\right)\right]^2$$

Least square estimation

Maximizing likelihood = minimizing error term

Linear Regression

$$\boxed{E(\theta \mid \mathcal{X}) = \frac{1}{2}\sum_{t=1}^{N}\left[r^t - g\left(x^t \mid \theta\right)\right]^2}$$

This is our loss function

$$\boxed{g\left(x^t \mid w_1, w_0\right) = w_1 x^t + w_0}$$

$$\underline{\sum_t r^t = N w_0 + w_1 \sum_t x^t}$$

$$\underline{\sum_t r^t x^t = w_0 \sum_t x^t + w_1 \sum_t \left(x^t\right)^2}$$

$$\mathbf{A} = \begin{bmatrix} N & \sum_t x^t \\ \sum_t x^t & \sum_t \left(x^t\right)^2 \end{bmatrix} \quad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} \sum_t r^t \\ \sum_t r^t x^t \end{bmatrix}$$

$$\mathbf{w} = \mathbf{A}^{-1}\mathbf{y}$$

Other Error Measures

- Square Error: $E(\theta \mid \mathcal{X}) = \frac{1}{2}\sum_{t=1}^{N}\left[r^t - g\left(x^t \mid \theta\right)\right]^2$
- Relative Square Error: $E\left(\theta \mid \mathcal{X}\right) = \frac{\sum_{t=1}^{N}\left[r^t - g\left(x^t \mid \theta\right)\right]^2}{\sum_{t=1}^{N}\left[r^t - \bar{r}\right]^2}$

If it is close to 1, it is almost like guessing the average of the data at all time

- Absolute Error: $E\,(\vartheta \mid \mathrm{X}) = \sum_t \left| r^t - g(x^t \mid \vartheta) \right|$

$$E\,(\vartheta \mid \mathrm{X}) = \sum_t 1(\left| r^t - g(x^t \mid \vartheta) \right| > \varepsilon)\,(\left| r^t - g(x^t \mid \vartheta) \right| - \varepsilon)$$

Error always decreases with more complex model
So how do we choose and select model?

Bias and Variance for regression

Given a dataset

No g, just pure noise
Variance of r

Difference between real output and our estimate output
Quantify how well on average our g(x) is on the training set

$$E\left[(r-g(x))^2 \mid x\right] = \underbrace{\boxed{E\left[(r-E[r \mid x])^2 \mid x\right]}}_{\textit{noise}} + \underbrace{\boxed{(E[r \mid x]-g(x))^2}}_{\textit{squared error}}$$

$$E_X\left[(E[r \mid x]-g(x))^2 \mid x\right] = \underbrace{\boxed{(E[r \mid x]-E_X[g(x)])^2}}_{\textit{bias}} + \underbrace{\boxed{E_X\left[(g(x)-E_X[g(x)])^2\right]}}_{\textit{variance}}$$

Average over dataset

To quantify how well on average our g(x) is on different dataset, we take the average over X

Side note:
Var(x) = E(X^2)-E(X)^2
Var(x) = E((x-E(x))^2)

Estimating Bias and Variance

- *Generate M* datasets of sampels X$_i$={x^t_i , r^t_i}, i=1,...,M to fit $g_i\,(x)$, i =1,...,M

$$\bar{g}(x) = \frac{1}{M}\sum_{i=1} g_i(X)$$

$$\text{Bias}^2(g) = \frac{1}{N}\sum_t \left[\bar{g}\left(x^t\right) - f\left(x^t\right)\right]^2$$

$$\text{Variance}(g) = \frac{1}{NM}\sum_t\sum_i \left[g_i\left(x^t\right) - \bar{g}\left(x^t\right)\right]^2$$

$$\bar{g}(x) = \frac{1}{M}\sum_t g_i(x)$$

Bias/Variance Dilemma

- Example: $g_i(x)$=2 has no variance and high bias

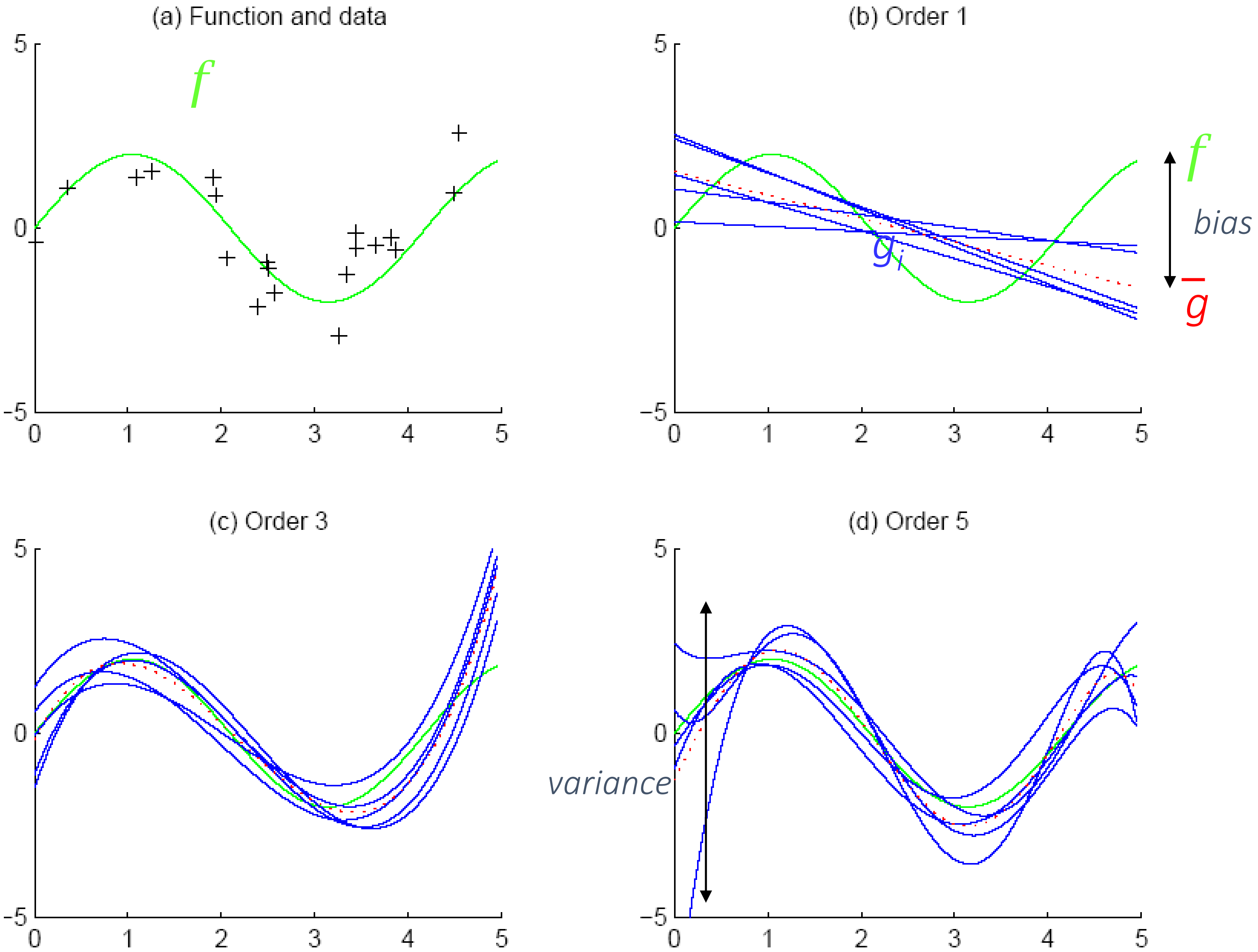
  $g_i(x)= \sum_t r^t_i/N$ has lower bias with variance

- As we increase complexity,

  bias decreases (a better fit to data) and

  variance increases (fit varies more with data)

- Bias/Variance dilemma: (Geman et al., 1992)


(a) Function and data
f
(b) Order 1
f
bias
g_i
$\bar{g}$
(c) Order 3
(d) Order 5
variance
5
0
-5
0
1
2
3
4
5

CHAPTER 5:

Multivariate Methods

Multivariate Data

- Multiple measurements (sensors)
- *d* inputs/features/attributes: *d*-variate
- *N* instances/observations/examples

$$\mathbf{X} = \begin{bmatrix} X_1^1 & X_2^1 & \cdots & X_d^1 \\ X_1^2 & X_2^2 & \cdots & X_d^2 \\ \vdots & & & \\ X_1^N & X_2^N & \cdots & X_d^N \end{bmatrix}$$

Multivariate Parameters

Multivariate Gaussian Distribution

$$\text{Mean}: E[\mathbf{x}] = \boldsymbol{\mu} = [\mu_1, \ldots, \mu_d]^T$$

$$\text{Covariance}: \sigma_{ij} \equiv \text{Cov}(X_i, X_j)$$

$$\text{Correlation}: \text{Corr}(X_i, X_j) \equiv \rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

$$\Sigma \equiv \text{Cov}(\mathbf{X}) = E\left[(\mathbf{X} - \mu)(\mathbf{X} - \mu)^T\right] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & & & \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{bmatrix}$$

Parameter Estimation

Sample mean $\mathbf{m}$: $m_i = \frac{\sum_{t=1}^{N} x_i^t}{N}, i = 1, \ldots, d$

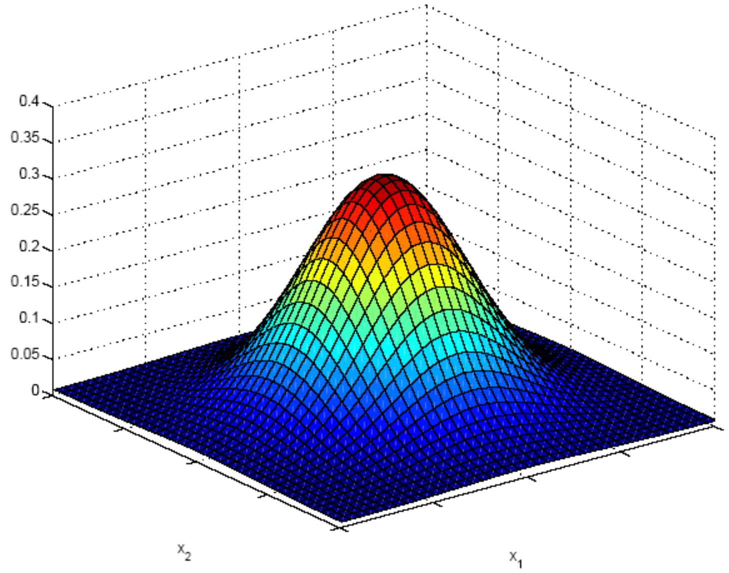
Covariance matrix $\mathbf{S}$: $s_{ij} = \frac{\sum_{t=1}^{N} \left(x_i^t - m_i\right)\left(x_j^t - m_j\right)}{N}$

Correlation matrix $\mathbf{R}$: $r_{ij} = \frac{s_{ij}}{s_i s_j}$

Estimation of Missing Values

- What to do if certain instances have missing attributes?
- Ignore those instances: not a good idea if the sample is small
- Use ‘missing’ as an attribute: may give information
- Imputation: Fill in the missing value
  - Mean imputation: Use the most likely value (e.g., mean)
  - Imputation by regression: Predict based on other attributes

Multivariate Normal Distribution



$$\mathbf{x} \sim \mathcal{N}_d(\boldsymbol{\mu}, \Sigma)$$

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

Multivariate Normal Distribution

Use of inverse variance

- Larger variance adds less distance
- Correlated variable contribute less

- Mahalanobis distance: $(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})$

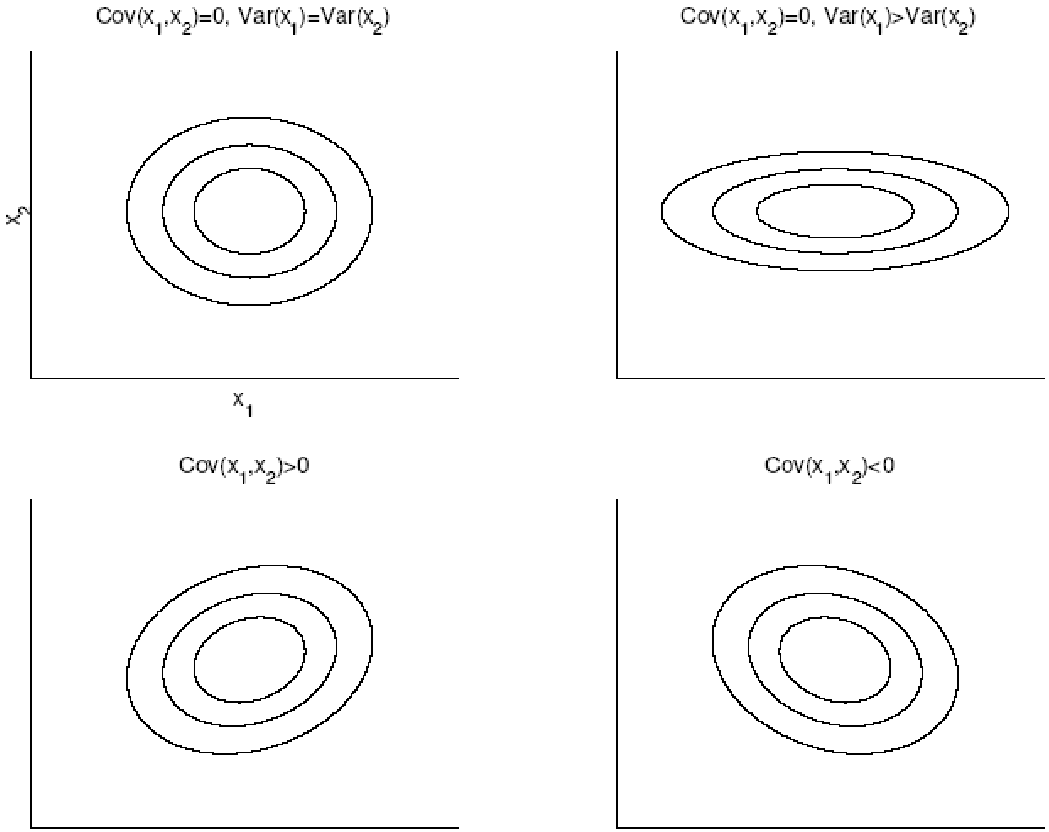
  measures the distance from $\boldsymbol{x}$ to $\boldsymbol{\mu}$ in terms of ∑ (normalizes for difference in variances and correlations)

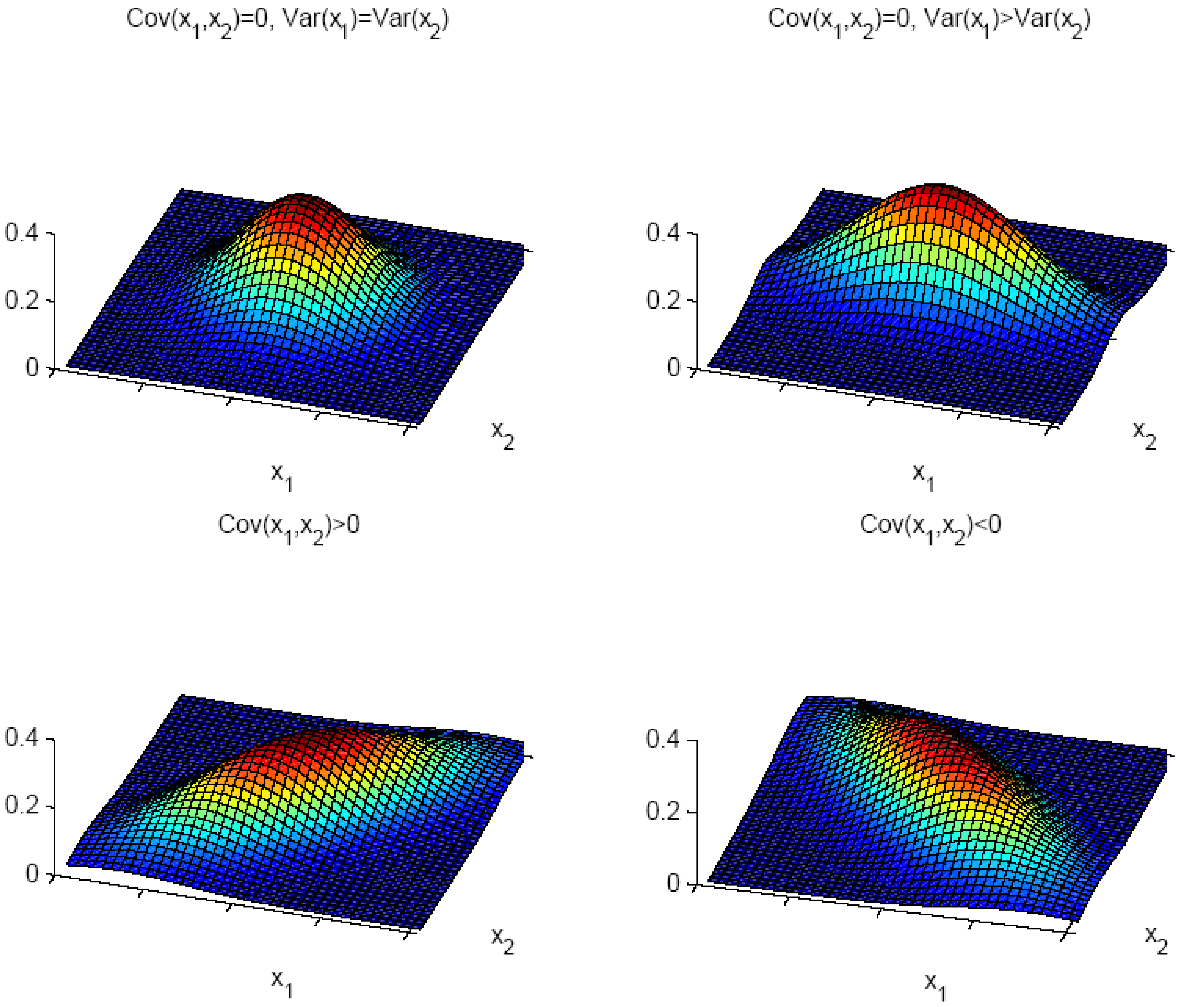
- Bivariate: $d = 2$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

$$\boxed{\begin{gathered} p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)} \left(z_1^2 - 2\rho z_1 z_2 + z_2^2 \right) \right] \\ z_i = (x_i - \mu_i)/\sigma_i \end{gathered}}$$

Bivariate Normal





$\mathrm{Cov}(x_1,x_2)=0$, $\mathrm{Var}(x_1)=\mathrm{Var}(x_2)$

$\mathrm{Cov}(x_1,x_2)=0$, $\mathrm{Var}(x_1)>\mathrm{Var}(x_2)$

$\mathrm{Cov}(x_1,x_2)>0$

$\mathrm{Cov}(x_1,x_2)<0$

Parametric Classification

- If $p\ (\boldsymbol{x} \mid C_i) \sim \mathrm{N}\ (\boldsymbol{\mu}_i , \Sigma_i)$

$$\boxed{p(\mathbf{x} \mid C_i) = \frac{1}{(2\pi)^{d/2} \left|\Sigma_i\right|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \right]}$$

- Discriminant functions

$$\boxed{\begin{aligned} g_i(\mathbf{x}) &= \log p(\mathbf{x} \mid C_i) + \log P(C_i) \\ &= -\frac{d}{2} \log 2\pi - \frac{1}{2} \log \left|\Sigma_i\right| - \frac{1}{2} (\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i) + \log P(C_i) \end{aligned}}$$

Estimation of Parameters

$$\hat{P}(C_i) = \frac{\sum_t r_i^t}{N}$$

$$\mathbf{m}_i = \frac{\sum_t r_i^t \mathbf{x}^t}{\sum_t r_i^t}$$

$$\mathbf{S}_i = \frac{\sum_t r_i^t (\mathbf{x}^t - \mathbf{m}_i)(\mathbf{x}^t - \mathbf{m}_i)^T}{\sum_t r_i^t}$$

$$g_i(\mathbf{x}) = -\frac{1}{2}\log|\mathbf{S}_i| - \frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T \mathbf{S}_i^{-1}(\mathbf{x}-\mathbf{m}_i) + \log \hat{P}(C_i)$$

Different S_i

Quadratic form

Quadratic discriminant

$$g_i(\mathbf{x}) = -\frac{1}{2}\log|\mathbf{S}_i| - \frac{1}{2}\left(\mathbf{x}^T\mathbf{S}_i^{-1}\mathbf{x} - 2\mathbf{x}^T\mathbf{S}_i^{-1}\mathbf{m}_i + \mathbf{m}_i^T\mathbf{S}_i^{-1}\mathbf{m}_i\right) + \log\hat{P}(C_i)$$

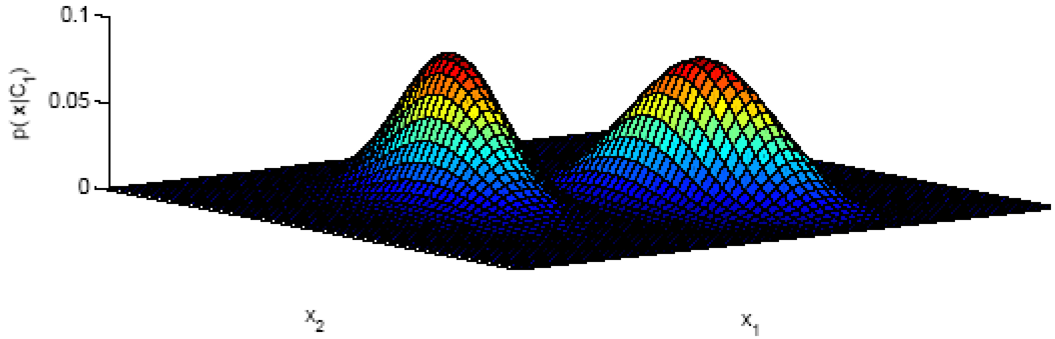
$$\boxed{= \mathbf{x}^T\mathbf{W}_i\mathbf{x} + \mathbf{w}_i^T\mathbf{x} + w_{i0}}$$

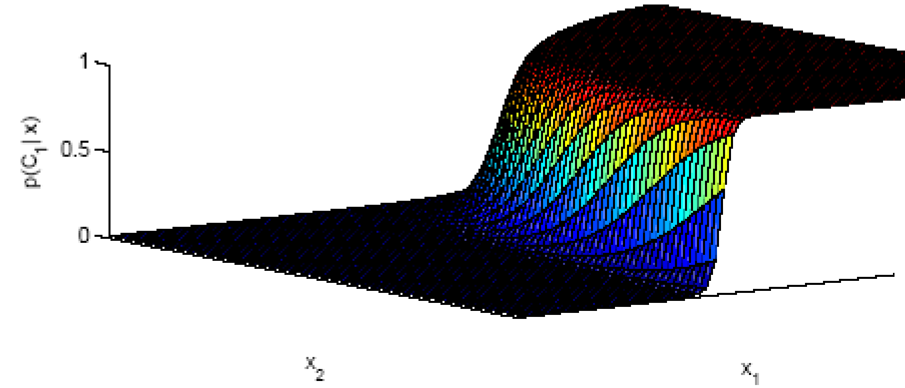
where

$$\mathbf{W}_i = -\frac{1}{2}\mathbf{S}_i^{-1}$$

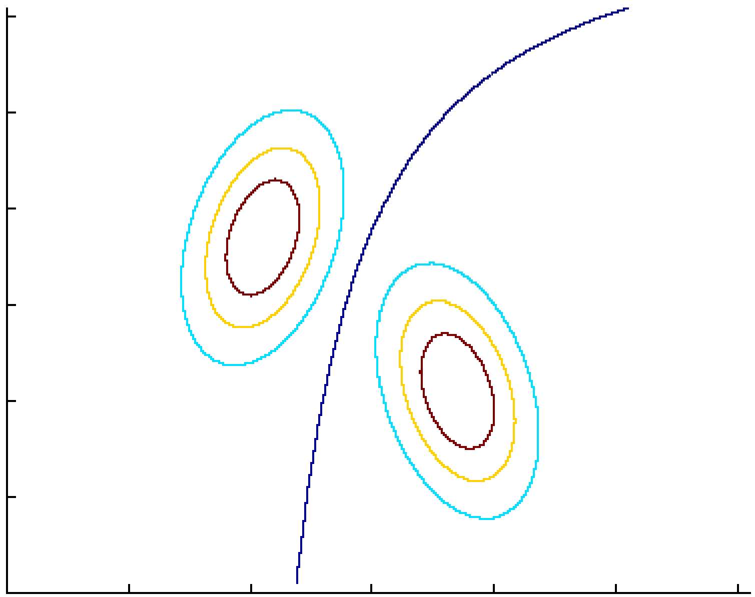
$$\mathbf{w}_i = \mathbf{S}_i^{-1}\mathbf{m}_i$$

$$w_{i0} = -\frac{1}{2}\mathbf{m}_i^T\mathbf{S}_i^{-1}\mathbf{m}_i - \frac{1}{2}\log|\mathbf{S}_i| + \log\hat{P}(C_i)$$





discriminant:
$P\,(C_1|\boldsymbol{x}) = 0.5$



Common Covariance Matrix **S**

- Shared common sample covariance **S** for all class

$$\mathbf{S} = \sum_i \hat{P}(C_i)\mathbf{S}_i$$

- Discriminant reduces to

$$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{m}_i)^T \mathbf{S}^{-1}(\boldsymbol{x} - \boldsymbol{m}_i) + \log \hat{P}(C_i)$$

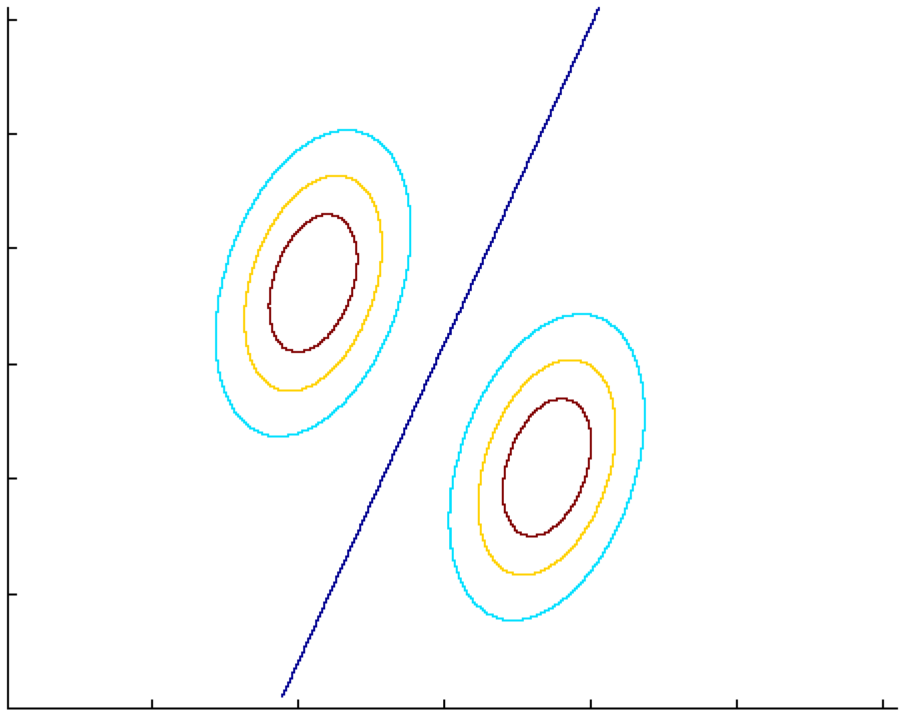
which is a linear discriminant

$$g_i(\mathbf{x}) = \mathbf{w}_i{}^T \mathbf{x} + w_{i0}$$

where

$$\mathbf{w}_i = \mathbf{S}^{-1}\mathbf{m}_i \quad w_{i0} = -\frac{1}{2}\mathbf{m}_i{}^T \mathbf{S}^{-1}\mathbf{m}_i + \log \hat{P}(C_i)$$

Common Covariance Matrix **S**



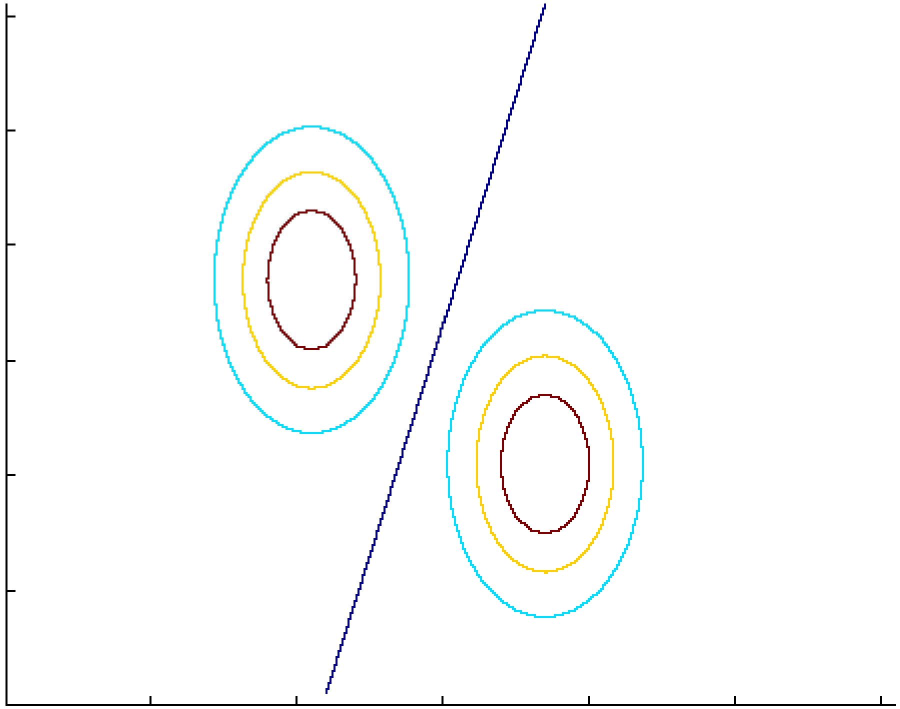
Diagonal **S**

- When x_j $j = 1,..d$, are independent, ∑ is diagonal
  $p\ (x|C_i) = \prod_j p\ (x_j\ |C_i)$ (Naive Bayes' assumption)

$$g_i(\mathbf{x}) = -\frac{1}{2}\sum_{j=1}^{d}\left(\frac{\boxed{x_j^t - m_{ij}}}{s_j}\right)^2 + \log \hat{P}(C_i)$$

Classify based on weighted Euclidean distance (in s_j units) to the nearest mean

Diagonal **S**



variances may be different

Independent Inputs: Naive Bayes

- If x_i are independent, off diagonals of ∑ are 0, Mahalanobis distance reduces to weighted (by 1/σ_i) Euclidean distance:

$$\boxed{p(\mathbf{x}) = \prod_{i=1}^{d} p_i(x_i) = \frac{1}{(2\pi)^{d/2} \prod_{i=1}^{d} \sigma_i} \exp\left[-\frac{1}{2}\sum_{i=1}^{d}\left(\frac{x_i - \mu_i}{\sigma_i}\right)^2\right]}$$

- If variances are also equal, reduces to Euclidean distance

Diagonal **S**, equal variances

- Nearest mean classifier: Classify based on Euclidean distance to the nearest mean

$$\boxed{\begin{aligned} g_i(\mathbf{x}) &= -\frac{\left\|\mathbf{x}-\mathbf{m}_i\right\|^2}{2s^2} + \log \hat{P}(C_i) \\ &= -\frac{1}{2s^2}\sum_{j=1}^{d}\left(x_j^t - m_{ij}\right)^2 + \log \hat{P}(C_i) \end{aligned}}$$

- Each mean can be considered a prototype or template and this is template matching

Diagonal **S**, equal variances