CS 559: Machine Learning: Fundamentals & Applications

Week 5: Linear Classification II

Naïve Bayes Classifier & Logistic Regression

Spring 24

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

4.5. Probability Theory – MLE vs. MAP
4.6. Naïve Bayes Classifier
4.7. Logistic Regression
4.8. Model Selection - Regularization

4.5. Probability Theory – MLE vs. MAP

4.6. Naïve Bayes Classifier

4.7. Logistic Regression

4.8. Model Selection - Regularization

4.5. Probability Theory – MLE vs. MAP

- Maximum Likelihood Estimator (MLE)
  - Estimates the parameter, θ, of the possible probability that describes the data.
  - $\boldsymbol{x} \sim p(\boldsymbol{x}|\theta)$
- Maximum A Posteriori Estimation (MAP)
  - Treat the parameter, θ, as a random variable that describes the data.
  - $p(\theta|\boldsymbol{x}) \propto p(\boldsymbol{x}|\theta)p(\theta)$

assumptions

↑ posterior prob.

likelihood ↳ prior prob.

$\theta \uparrow \Rightarrow$ Error ↓

4.5. Probability Theory – MLE vs. MAP

- Example Coin Toss
  - Suppose a fair coin is tossed N times and each toss result is recorded.
- Given data set:
  - $\boldsymbol{x} = \{x_i\}_{i=1}^{N}, y_i \in \{-1, 1\}$
  - -1 is the tail and 1 is the head.
- How can θ be estimated given **x**?

$$\hat{\theta} = \frac{\sum_{i=1}^{N} \boldsymbol{I}(y_i = 1)}{N}$$

count # of heads

total # of toss

4.5. Probability Theory – MLE vs. MAP

- MLE Approach
- Let's assume the probability of y is in the Bernoulli distribution.
- The Bernoulli distribution is the discrete probability distribution
  - p(y=1) = p ← head
  - p(y=0) = 1-p ← tail
  - $f(k|p) = p^k(1-p)^{1-k}$

4.5. Probability Theory – MLE vs. MAP

- The example likelihood can be expressed as

$$p(x|\theta) = \theta^{n_1}(1-\theta)^{N-n_1}$$

$$y \sim p(x|\theta)$$

where n_1 is the count of y=1.

- Estimate $\hat{\theta}_{MLE}$: $\hat{\theta}_{MLE} = \frac{n_1}{N}$

$\hat{\theta}_{MLE} = \arg\max_{\theta} (P(X|\theta)) = \arg\max_{\theta} \left(\theta^{n_1}(1-\theta)^{N-n_1}\right)$

$= \arg\max_{\theta} \ln\left(\theta^{n_1}(1-\theta)^{N-n_1}\right)$

$= \arg\max_{\theta} \left[n_1 \ln\theta + (N-n_1)\ln(1-\theta)\right]$

n_1 = # of $y=1$ (head)

N = ~~total~~ # of obs. (toss)

$n_1 = \frac{\sum \mathbb{I}(y=1)}{N}$

θ^{n_1} = total prob. of head

$(1-\theta)^{N-n_1}$ = total prob of tail.

$\ln(ab) = \ln a + \ln b$

$\ln(b^x) = x \ln b$

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \underbrace{\left(n_1 \ln \theta + (N - n_1) \ln (1 - \theta) \right)}$$

take a derivative w.r.t. θ & set $= 0$.

$$= \arg\max_{\theta} \left(\frac{n_1}{\theta} - \frac{N - n_1}{1 - \theta} \right)$$

$= 0$

$$\frac{n_1}{\theta} = \frac{N - n_1}{1 - \theta} \Rightarrow \hat{\theta}_{MLE} = \frac{n_1}{N}$$

$f(\theta) = \ln \theta$

$f'(\theta) = \frac{df}{d\theta} = \frac{1}{\theta}$

$\frac{d}{d\theta} \left(\ln(1 - \theta) \right)$

$= \frac{1}{1 - \theta} \frac{d}{d\theta} (1 - \theta)$

$= -\frac{1}{1 - \theta}$

4.5. Probability Theory – MLE vs. MAP

- MAP approach
- Let θ is a random variable and estimate the posterior probability of θ given $\mathbf{x}$.
- Use the Bayes rule:

$$\text{Posterior} \propto \text{likelihood} * \text{prior}$$

$$p(\theta|x) \propto p(x|\theta)p(\theta)$$

learning ← assumption

- The data set is $\boldsymbol{x} = \{y_i\}_{i=1}^{N}$.
- The actual Bayes rule is

$$p(\theta|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{p(\boldsymbol{x})}$$

← actual distribution of data very hard to know

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\theta)p(\theta)\,\mathrm{d}\theta$$

4.5. Probability Theory – MLE vs. MAP

Steps to calculate $\hat{\theta}_{MAP}$:

- Prior probability: assume $p(\theta) \sim \theta^{\alpha-1}(1-\theta)^{\beta-1}$
- Likelihood: Bernoulli $p(x|\theta)$ [known] N + n_1 $\theta^{n_1}(1-\theta)^{N-n_1} = P(x|\theta)$
- Posterior: compute $p(\theta)p(\boldsymbol{x}|\theta)$
- Maximize the posterior.
- Solution of the example: $\hat{\theta}_{MAP} = \frac{(n_1+\alpha-1)}{N+\alpha+\beta+2}$

$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}} \left(\theta^{n_1}(1-\theta)^{N-n_1} \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1}\right)$

1. log- posterior
2. $\frac{\partial}{\partial\theta}$(log- posterior) $= 0$
3. Solve for θ.

$$P(\theta|x) = \left[\theta^{n_1}(1-\theta)^{N-n_1}\right]\left[\theta^{\alpha-1}(1-\theta)^{\beta-1}\right]$$

$$\ln(P(\theta|x)) = n_1 \ln\theta + (N-n_1)\ln(1-\theta) + (\alpha-1)\ln\theta + (\beta-1)\ln(1-\theta)$$

$$= (n_1+\alpha-1)\ln\theta + (N-n_1+\beta-1)\ln(1-\theta)$$

$$\frac{\partial}{\partial\theta}\ln(P(\theta|x)) = \frac{n_1+\alpha-1}{\theta} - \frac{(N-n_1+\beta-1)}{1-\theta} = 0$$

$$\frac{\theta}{1-\theta} = \frac{n_1+\alpha-1}{N-n_1+\beta-1}$$

if $N \to \infty$, $\lim_{N\to\infty}\left(\frac{n_1+\alpha-1}{N-n_1+\beta-1}\right) \simeq \frac{n_1}{N-n_1} \simeq \frac{\theta}{1-\theta} \Rightarrow \hat{\theta}_{map} = \frac{n_1}{N}$

MLE

4.5. Probability Theory – MLE vs. MAP

- Example: Continuous $\mathbf{x} = \{x_i\}_{i=1}^{N}, x_i \in \mathbb{R}^d$.
- Assume each x is independent.
- The point distribution assumption:

$$x \sim \mathcal{N}(x|\mu, \sigma^2 = 1)$$
$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x-\mu)^2\right)$$

- Then the data distribution becomes

$$p(x|\mu) = \prod_{i=1}^{N} \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{1}{2}(x_i-\mu)^T(x_i-\mu)\right).$$

- The solution is $\hat{\mu}_{MLE} = \frac{\sum_{i=1}^{N} x_i}{N}$.

4.5. Probability Theory – MLE vs. MAP

- MLE vs. MAP:
- MAP works better when N is small, and the prior distribution is known.
- However, knowing the prior distribution for a data set with finite examples is difficult.
- MLE is preferred for large N.

4.5. Probability Theory – MLE vs. MAP

4.6. Naïve Bayes Classifier

4.7. Logistic Regression

4.8. Model Selection - Regularization

4.6. Naïve Bayes Classifier

posterior $\propto p(x|\omega)P(\omega)$

- The Naïve Bayes classifier is based on applying Bayes' theorem.
- Assumptions: Features are **conditionally independent** of a given target.
- Pros: Easy to train and predict fast. Sometimes, it is better than other linear classifiers.
- Cons: Strong assumptions.
  - It is not realistic in real life.
  - Computation becomes complex if training and test sets have different categories (We use Laplace estimation, see Bishop 4.4).

4.6. Naïve Bayes Classifier

- From Bayes' theorem, the conditional probability (posterior) can be decomposed as

$$p(C_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_k)p(C_k)}{p(\boldsymbol{x})}$$

  where C_k is the class and $\boldsymbol{x}$ is the features.

- Remember that we are only interested in the numerator (=likelihood × prior).
- The numerator is equivalent to the joint probability model using the chain rule.

4.6. Naïve Bayes Classifier

- When two variables, A and B, are independent,

$$p(A \perp B) = p(A)p(B)$$

$$\forall\, a, b: p(A = a \perp B = b) = p(A = a)p(B = b).$$

- If two variables, A and B, are **conditionally independent** given C,

$$p(A, B|C) = p(A|C)p(B|C)$$

$$\forall\, a, b, c: p(A = a \perp B = b|C = c) = p(A = a|C = c)p(B = b|C = c).$$

4.6. Naïve Bayes Classifier

- Suppose there are D-many features, $\boldsymbol{x} = [x_1, \dots, x_D]$.
- The joint probability is $p(C_k, x_1, \dots, x_n)$.
- Using the chain rule for the definition of conditional probability:

$$p(C_k, x_1, \dots, x_n) = p(x_1, \dots, x_D, C_k)$$
$$= p(x_1 | x_2, \dots, C_k) p(x_2 | x_3, \dots, C_k) \cdots p(x_D | C_k) p(C_k)$$

4.6. Naïve Bayes Classifier

- Under the conditional independence assumption, the joint model can be expressed as

$$p(C_k|x_1, \dots, x_n) \propto p(C_k, x_1, \dots, x_n)$$
$$\propto p(C_k)p(x_1|C_k)p(x_2|C_k)\cdots$$
$$\propto p(C_k)\prod_{i=1}^{n} p(x_i|C_k).$$

prior ← likelihood

- The denominator $p(\boldsymbol{x})$ is

$$p(\boldsymbol{x}) = \sum_k p(C_k)p(\boldsymbol{x}|C_k).$$

4.6. Naïve Bayes Classifier

- Finding $p(C_k)$ is easy:

$$p(C_k = k) = \frac{\sum_{i=1}^{N} I(C_k = k)}{N} = \pi_k$$

- The likelihood is

$$p(x|C_k) = \prod_{i=1}^{D} p(x_i|C_k)$$

$$p(C_k|x) \propto p(C_k) \prod_{i=1}^{D} p(x_i|C_k)$$

$$\ln(p(C_k|x)) = \ln p(C_k) + \ln\left(\prod_{i=1}^{D} p(x_i|C_k)\right)$$

4.6. Naïve Bayes Classifier

- The Bayes classifier is then

$$h(x) = \underset{C_k}{\operatorname{argmax}} \left(\sum_{i=1}^{D} \ln\left(p(\boldsymbol{x}|C_k)\right) + \ln\left(p(C_k)\right) \right)$$

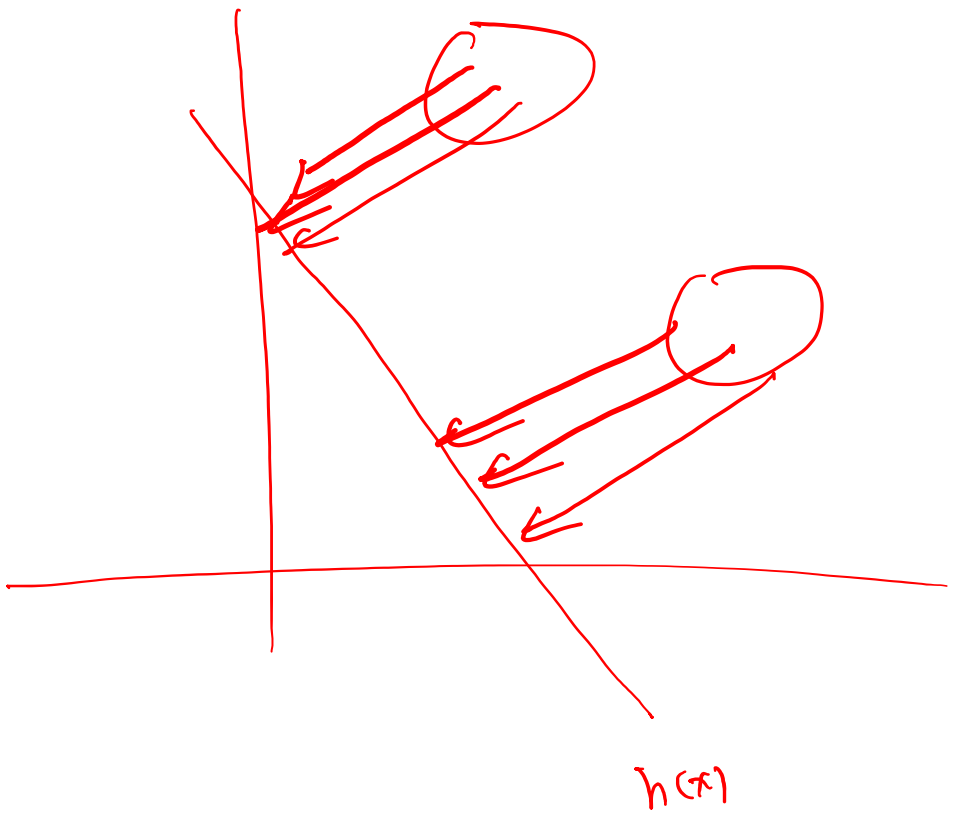
??

$h(\vec{x}) = \omega^T \vec{x} + \omega_0$

4.6. Naïve Bayes Classifier

Assume features hold the LDA conditions. $\Sigma_1 = \Sigma_2 = \Sigma$

The Bayes classifier will show that the model is a linear equation.



$$\vec{\omega} \propto S_W^{-1}(\vec{m}_2 - \vec{m}_1)$$

$$\vec{\omega} = \Sigma^{-1}(\vec{\mu}_2 - \vec{\mu}_1)$$

4.6. Naïve Bayes Classifier

- Consider a binary class $y_i \in \{-1, 1\}$ and the continuous feature $\boldsymbol{x} \in \mathbb{R}^D$.
- Let the prior probability be $p(C_1) = \pi_1$ and $p(C_2) = \pi_2$.
- Assume x is in Gaussian.
- The probability distribution function (PDF) of the likelihood for class C_k is

$$p(\boldsymbol{x}|C_k) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu_k})^T \boldsymbol{\Sigma^{-1}} (\boldsymbol{x} - \boldsymbol{\mu_k})\right\}.$$

4.6. Naïve Bayes Classifier

- The posterior becomes for C_1:

$$p(C_1|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_1)p(C_1)}{p(\boldsymbol{x}|C_1)p(C_1) + p(\boldsymbol{x}|C_2)p(C_2)}$$

- Solutions:

$$\boldsymbol{w^T} = (\boldsymbol{\mu_1^T} - \boldsymbol{\mu_2^T})\boldsymbol{\Sigma^{-1}}$$

$$w_0 = \frac{1}{2}(\boldsymbol{\mu_1^T\Sigma^{-1}\mu_1} - \boldsymbol{\mu_2^T\Sigma^{-1}\mu_2}) + \ln\frac{\pi_1}{\pi_2}$$

$$\text{post.} = \frac{P(C_k)\,p(\vec{x}\mid C_k)}{p(\vec{x})}$$

$P(C_1) = \pi_1 \quad P(C_2) = \pi_2$

$p(\vec{x}\mid C_1) = \text{const}_1 \cdot \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu}_1)^T \Sigma_1^{-1} (\vec{x}-\vec{\mu}_1)\right)$

$p(\vec{x}\mid C_2) = \text{const}_2 \cdot \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu}_2)^T \Sigma_2^{-1} (\vec{x}-\vec{\mu}_2)\right)$

$\Sigma_1 = \Sigma_2 = \Sigma$

$\text{const}_1 = \text{const}_2 = \text{same}$

$$\pi_1 \text{const}_1 \cdot \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu}_1)^T \Sigma_1^{-1} (\vec{x}-\vec{\mu}_1)\right)$$

$$\pi_1 \text{const}_1 \cdot \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu}_1)^T \Sigma_1^{-1} (\vec{x}-\vec{\mu}_1)\right) + \pi_2 \text{const}_2 \cdot \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu}_2)^T \Sigma_2^{-1} (\vec{x}-\vec{\mu}_2)\right)$$

$$A = -\frac{1}{2}(\vec{x}-\vec{\mu}_1)^T \Sigma^{-1} (\vec{x}-\vec{\mu}_1)$$

$$B = -\frac{1}{2}(\vec{x}-\vec{\mu}_2)^T \Sigma^{-1} (\vec{x}-\vec{\mu}_2)$$

$$\text{post} = \frac{\pi_1 \exp(A)}{\pi_1 \exp(A) + \pi_2 \exp(B)} = \frac{1}{1 + \frac{\pi_2}{\pi_1} \frac{\exp B}{\exp A}}$$

$$= \frac{1}{1 + \underbrace{\pi_2/\pi_1}\exp(B - A)}$$

$$= \frac{1}{1 + \exp\left(B - A + \ln \pi_2/\pi_1\right)}$$

$$\ln(\exp(a)) = a$$

$$A = -\frac{1}{2}(\vec{x}-\mu_1)^T \Sigma^{-1} (\vec{x}-\mu_2) = -\frac{1}{2}\left[\vec{x}^T\Sigma^{-1}\vec{x} - 2\vec{x}^T\Sigma^{-1}\vec{\mu}_1 + \vec{\mu}_1^T\Sigma^{-1}\vec{\mu}_1 \right]$$

$$B = -\frac{1}{2}\left(\vec{x}^T\Sigma^{-1}\vec{x} - 2\vec{x}^T\Sigma^{-1}\vec{\mu}_2 + \vec{\mu}_2^T\Sigma^{-1}\vec{\mu}_2 \right]$$

$$B-A = \vec{x}^T\Sigma^{-1}\vec{\mu}_2 - \frac{1}{2}\vec{\mu}_2\Sigma^{-1}\vec{\mu}_2 - \vec{x}^T\Sigma^{-1}\vec{\mu}_1 + \frac{1}{2}\vec{\mu}_1\Sigma^{-1}\vec{\mu}_1$$

$$= -\left(\Sigma^{-1}\vec{x}^T(\vec{\mu}_1 - \mu_2)\right) - \frac{1}{2}\left(\Sigma^{-1}(\vec{\mu}_2^T\vec{\mu}_2 - \vec{\mu}_1^T\vec{\mu}_1)\right)$$

$$\text{Post} = \frac{1}{1 + \exp(B - A + \ln \pi_2/\pi_1)}$$

$$B - A = -\Sigma^{-1}(\vec{\mu}_1 - \vec{\mu}_2)\vec{x}^T - \frac{1}{2}\Sigma^{-1}(\vec{\mu}_2^T\vec{\mu}_2 - \vec{\mu}_1^T\vec{\mu}_1)$$

$$\text{post} = \frac{1}{1 + \exp\left(-\Sigma^{-1}(\vec{\mu}_1 - \vec{\mu}_2)\vec{x}^T - \frac{1}{2}\Sigma^{-1}(\vec{\mu}_2^T\vec{\mu}_2 - \vec{\mu}_1^T\vec{\mu}_1) + \ln \pi_2/\pi_1\right)}$$

Σ^{-1}: const; $(\vec{\mu}_1 - \vec{\mu}_2)$: const; $\vec{x}^T$: Data; $\vec{w}$

$\frac{1}{2}\Sigma^{-1}(\vec{\mu}_2^T\vec{\mu}_2 - \vec{\mu}_1^T\vec{\mu}_1)$: const; $\ln \pi_2/\pi_1$: const.; w_0

$$= \frac{1}{1 + \exp\left(-\Sigma^{-1}(\vec{\mu}_1^T - \vec{\mu}_2^T)\vec{x} + \frac{1}{2}(\vec{\mu}_1^T\Sigma^{-1}\vec{\mu}_1 - \vec{\mu}_2^T\Sigma^{-1}\vec{\mu}_2) + \ln \pi_2/\pi_1\right)}$$

$$\text{post} = \frac{1}{1 + \exp(-(w^T x + w_0))}$$

$$w^T = (\vec{\mu}_1^T - \vec{\mu}_2^T)\Sigma^{-1}$$

$$w_0 = \frac{1}{2}(\vec{\mu}_1^T \Sigma^{-1} \vec{\mu}_1 - \vec{\mu}_2^T \Sigma^{-1} \mu_2) + \ln \pi_1/\pi_2$$

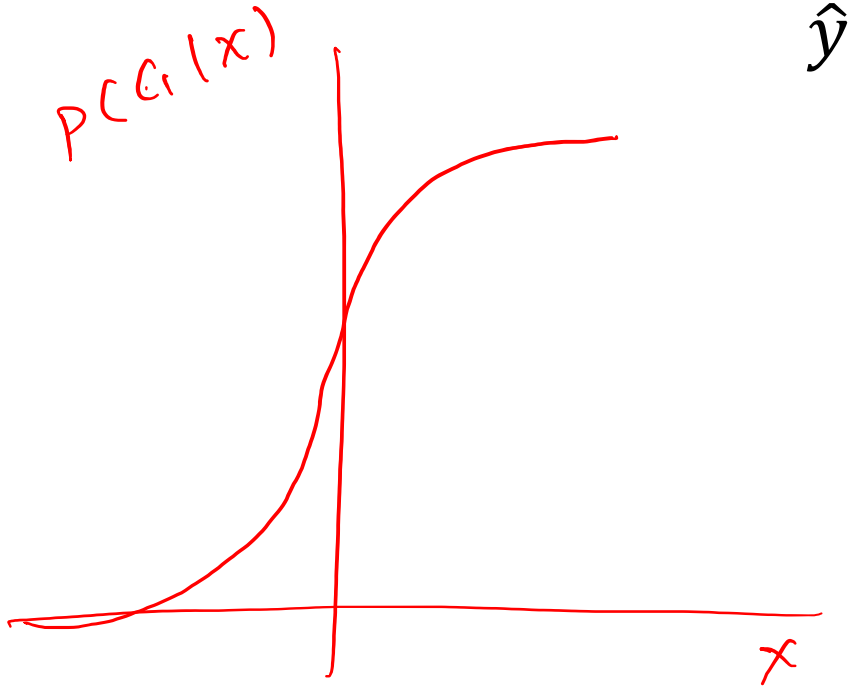
4.6. Naïve Bayes Classifier

- The solution, $\hat{h} = \boldsymbol{w}^T\boldsymbol{x} + \boldsymbol{w}_0$, and $p(C_1|\boldsymbol{x})$ is called a **sigmoid function**.

$$p(C_1|\boldsymbol{x}) = \frac{1}{1 + \exp(-\hat{h})}$$

- The predicted class for **x** is

$$\hat{y} = \begin{cases} -1 & \text{if } \hat{h} < 0.5 \\ 1 & \text{otherwise} \end{cases}$$



LDA : $h = w^T x + w_0$

$h < 0$ or $h \geq 0$

$C = -1$ $C = 1$

BC $\frac{1}{1+\exp(-h)} < 0.5$ or

prob. $C = -1$

≥ 0.5

$C = 1$

4.6. Naïve Bayes Classifier

- Sigmoid function

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

4.6. Naïve Bayes Classifier

- How about MLE?
- Assume **y** is in the Bernoulli distribution and **X** is continuous data and is in Gaussian distribution.
- Each class has a Gaussian class-conditional density with a shared $\boldsymbol{\Sigma}$.
- $y_n = \{1,0\}$ for C_1 and C_2, respectively.
- If the prior class probability is $p(C_1) = \pi$, $p(C_2) = 1 - \pi$.

4.6. Naïve Bayes Classifier

- For a data point $\boldsymbol{x}_i$ from class C_1, $y_i = 1$ and hence the joint probability for each class becomes

$$p(\boldsymbol{x}_i, C_1) = p(C_1)p(\boldsymbol{x}_i|C_1) = \pi\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu_1}, \boldsymbol{\Sigma})$$
$$p(\boldsymbol{x}_i, C_2) = p(C_2)p(\boldsymbol{x}_i|C_2) = (1-\pi)\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu_2}, \boldsymbol{\Sigma}).$$

- Then the likelihood function is given by

$$p(\boldsymbol{X}|\boldsymbol{y}) = \prod_{i=1}^{N} [\pi\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu_1}, \boldsymbol{\Sigma})]^{y_i}[(1-\pi)\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu_2}, \boldsymbol{\Sigma})]^{1-y_i}$$

$\theta^{n_1}(1-\theta)^{N-n_1}$

4.6. Naïve Bayes Classifier

- Solutions:

$$\pi = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{N_1}{N}$$

where N_1 is the number of data points in class C_1.

$$\boldsymbol{\mu_1} = \frac{1}{N_1}\sum_{i=1}^{N} y_i \boldsymbol{x}_i \,, \boldsymbol{\mu_2} = \frac{1}{N_2}\sum_{i=1}^{N} (1 - y_i)\boldsymbol{x}_i$$

$$\boldsymbol{\Sigma} = \frac{N_1}{N}\boldsymbol{S}_1 + \frac{N_2}{N}\boldsymbol{S}_2$$

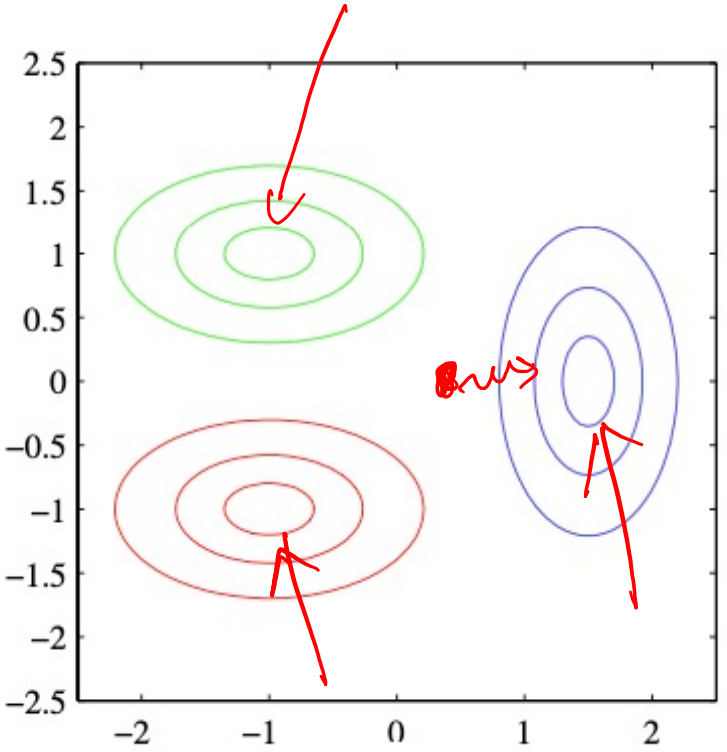
Recall, $\boldsymbol{w^T} = (\boldsymbol{\mu_1^T} - \boldsymbol{\mu_2^T})\boldsymbol{\Sigma^{-1}}$.

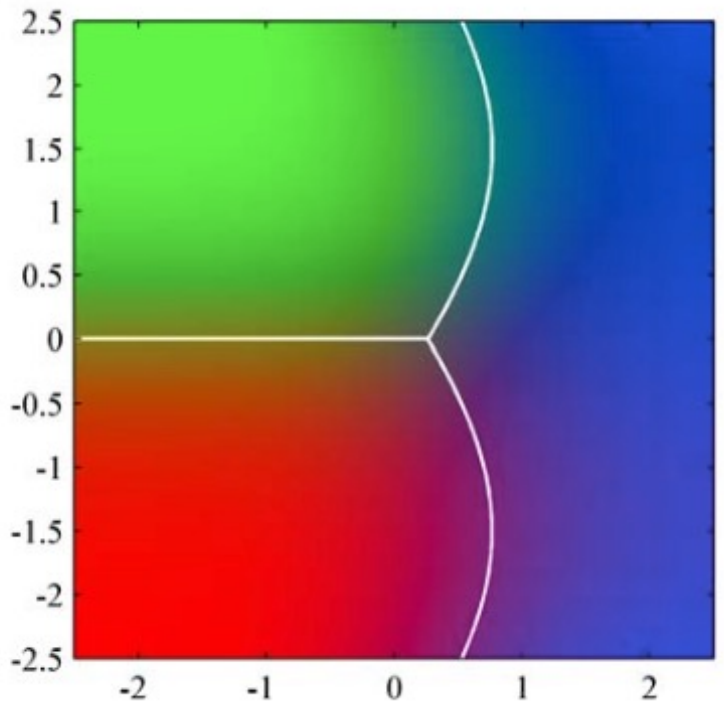
4.6. Naïve Bayes Classifier

- For the case of $K > 2$ classes, $\hat{h}(\boldsymbol{x}) = \boldsymbol{w}_k^T \boldsymbol{x} + w_{k0}$

  where $\boldsymbol{w_k} = \boldsymbol{\Sigma^{-1}\mu_k}$ and $w_{k0} = -\frac{1}{2}\boldsymbol{\mu}_k^T \boldsymbol{\Sigma^{-1}\mu}_k + \ln p(C_k)$

- If each class has its own covariance matrix $\boldsymbol{\Sigma}_k$, then $\boldsymbol{x\Sigma_k^{-1}x}$ term will not vanish.





4.6. Naïve Bayes Classifier

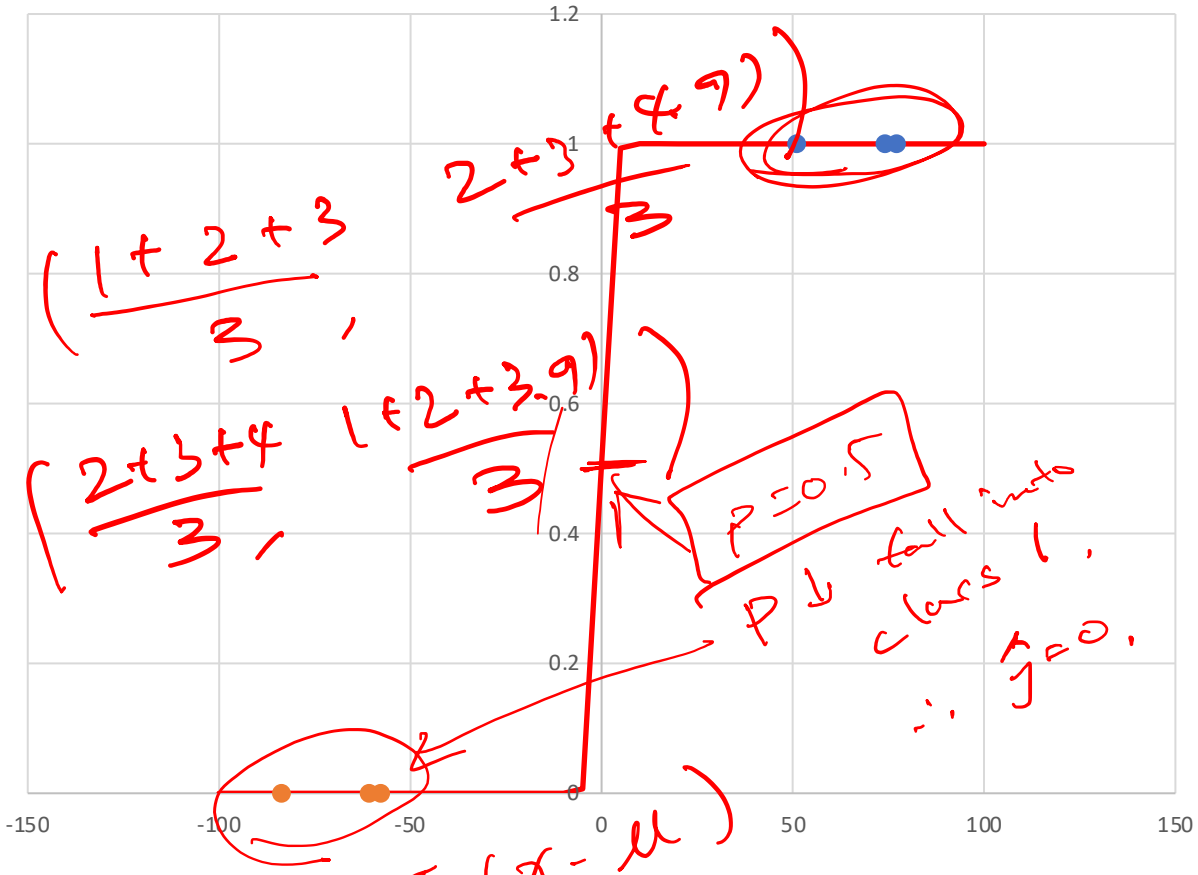
Class	(x_1, x_2)
1	(1,2), (2,3), (3,4.9)
0	(2,1), (3,2), (4, 3.9)

$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.667 & 0.967 \\ 0.967 & 1.467 \end{bmatrix}$$

$$\boldsymbol{\Sigma^{-1}} = \begin{bmatrix} 48.22 & -32.22 \\ -32.22 & 22.22 \end{bmatrix}$$

$$\boldsymbol{w} = (\boldsymbol{\mu_1^T} - \boldsymbol{\mu_2^T})\boldsymbol{\Sigma^{-1}} = [48.667, -80.444, 54.444]$$

$$\hat{\boldsymbol{w}} = \frac{\boldsymbol{w}}{|\boldsymbol{w}|} = [\mathbf{0.448}, \mathbf{-0.828}, \mathbf{0.560}]$$



$h = w^T x$

$y = \frac{1}{1+\exp(-h)}$

$y \geq 0.5, \hat{y} = 1$

$< 0.5, \hat{y} = 0$

$\sum (x-\mu)^T (x-\mu)$