Classification: Naïve Bayes, MLE, MAP wrap-up Logistic Regression

Aarti Singh

Co-instructor: Barnabas Poczos

Machine Learning 10-401
Jan 26, 2016

Announcements

- Recitation tomorrow 6:30-7:30 pm GHC 4303
  on Convex functions, linear algebra
  By Kirstin Early

- Project teams of 2
  competition of a given dataset

Classification so far …

$$X = (X_1, \ldots, X_d)$$
$$x = (x_1, \ldots, x_d)$$

- Bayes Optimal Classifier

$$f^*(x) = \arg\max_{Y=y} P(Y = y|X = x)$$

$$= \arg\max_{Y=y} P(X = x|Y = y)P(Y = y)$$

- Naïve Bayes Classifier

$$f_{NB}(x) = \arg\max_{Y=y} \prod_{i=1}^{d} P(X_i = x_i|Y = y)P(Y = y)$$

Assume parametric models for class probability and class conditional feature distribution and estimate parameters using MLE or MAP

Parametric Models

- For class probability

  Two classes: $P(Y=y) \sim$ Bernoulli

  Multiple classes: $P(Y=y) \sim$ Multinomial

- For class conditional feature distribution

  Binary features: $P(X_i = x_i | Y=y) \sim$ Bernoulli

  Discrete features: $P(X_i=x_i | Y=y) \sim$ Multinomial

  Continuous features: $P(X_i=x_i | Y=y) \sim$ Gaussian

Estimate parameters of distributions using MLE or MAP

MLE and MAP

- Maximum Likelihood estimation (MLE)

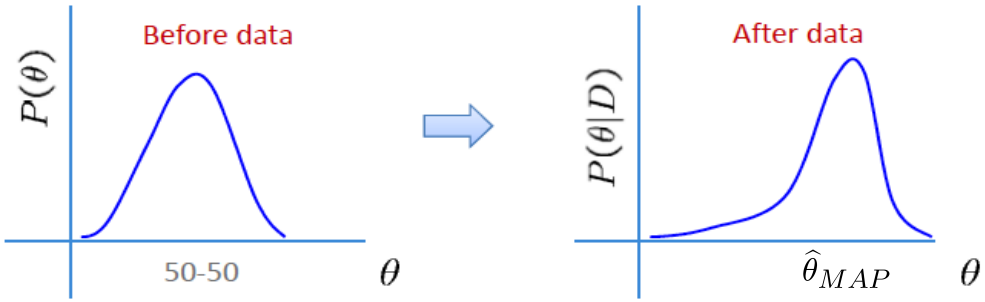
  Choose value that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D|\theta)$$

- Maximum *a posteriori* (MAP) estimation

  Choose value that is most probable given observed data and prior belief

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta|D)$$
$$= \arg\max_{\theta} P(D|\theta)P(\theta)$$



MLE and MAP

- Bernoulli(θ)

Beta(β_H, β_T) prior

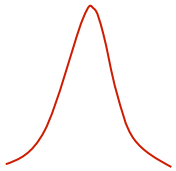
$$\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T} \qquad \hat{\theta}_{MAP} = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2}$$

- Multinomial(p_1, p_2, …, p_K)

Dirichlet(β_1,…, β_K) prior

$$\hat{p}_{y,MLE} = \frac{\alpha_y}{\sum_y \alpha_y} \qquad \hat{p}_{y,MAP} = \frac{\alpha_y + \beta_y - 1}{\sum_y (\alpha_y + \beta_y - 1)}$$

- Gaussian(μ,σ)



$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

Gaussian prior for mean
Wishart prior for variance
(not required)

Naïve Bayes Algo – Discrete features

- Training Data $\{(X^{(j)}, Y^{(j)})\}_{j=1}^{n}$ $X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)})$
- Maximum Likelihood Estimates
  - For Class probability $\hat{P}(y) = \frac{\{\#j : Y^{(j)} = y\}}{n}$
  - For class conditional feature distribution

$$\hat{P}(x_i|y) = \frac{\hat{P}(x_i, y)}{\hat{P}(y)} = \frac{\{\#j : X_i^{(j)} = x_i, Y^{(j)} = y\}/n}{\{\#j : Y^{(j)} = y\}/n}$$

- NB Prediction for test data $X = (x_1, \ldots, x_d)$

$$Y = \arg\max_y \hat{P}(y) \prod_{i=1}^{d} \frac{\hat{P}(x_i, y)}{\hat{P}(y)}$$

Naïve Bayes Algo – Discrete features

- Training Data $\{(X^{(j)}, Y^{(j)})\}_{j=1}^{n}$ $X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)})$
- MLE for Class probability, MAP for Class conditional feature dist
  - For Class probability $$\hat{P}(y) = \frac{\{\#j : Y^{(j)} = y\}}{n}$$
  - For class conditional feature distribution add virtual examples $$\hat{P}(x_i|y) = \frac{\{\#j : X_i^{(j)} = a, Y^{(j)} = y\} + \beta_a^{(y)} - 1}{\{\#j : Y^{(j)} = y\} + \sum_a (\beta_a^{(y)} - 1)}$$
- NB Prediction for test data $X = (x_1, \ldots, x_d)$ $$Y = \arg\max_y \hat{P}(y) \prod_{i=1}^{d} \frac{\hat{P}(x_i, y)}{\hat{P}(y)}$$

Naïve Bayes Algo – continuous features

- Training Data $\{(X^{(j)}, Y^{(j)})\}_{j=1}^{n}$ $X^{(j)} = (X_1^{(j)}, \ldots, X_d^{(j)})$
- Maximum Likelihood Estimates
  - For Class probability $\hat{P}(y) = \frac{\{\#j : Y^{(j)} = y\}}{n}$
  - For class conditional distribution

$$\hat{P}(x_i|y) = N(\hat{\mu_i}^{(y)}, \hat{\sigma_i}^{2(y)})$$

MLE estimates

- NB Prediction for test data $X = (x_1, \ldots, x_d)$

$$Y = \arg\max_y \hat{P}(y) \prod_{i=1}^{d} \hat{P}(x_i|y)$$

Applications of Naïve Bayes

- Text classification

  Documents ➔ Topic

  Emails ➔ Spam

  HW1

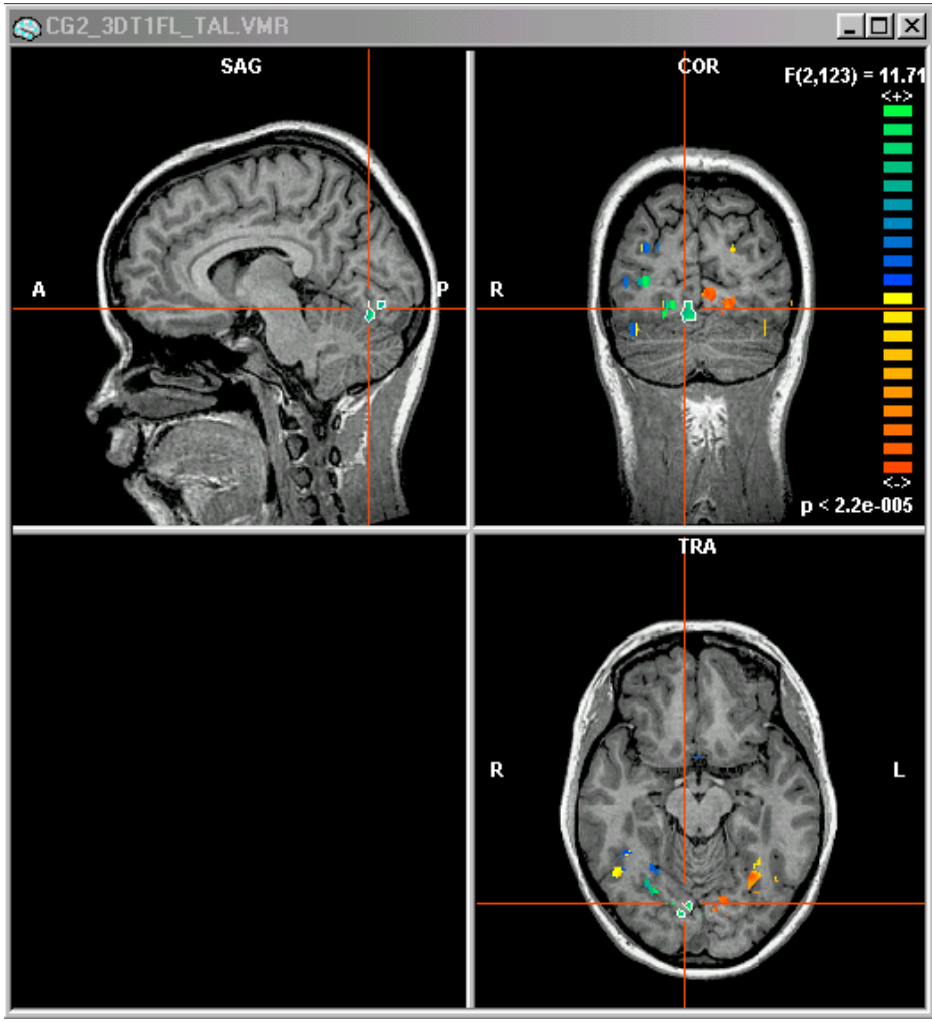
- Image classification

  Hand-written characters ➔ Digit, Letter

  Brain scans ➔ words, objects person is reading/ watching

GNB for classifying word category

[Mitchell et al.03]



~1 mm resolution

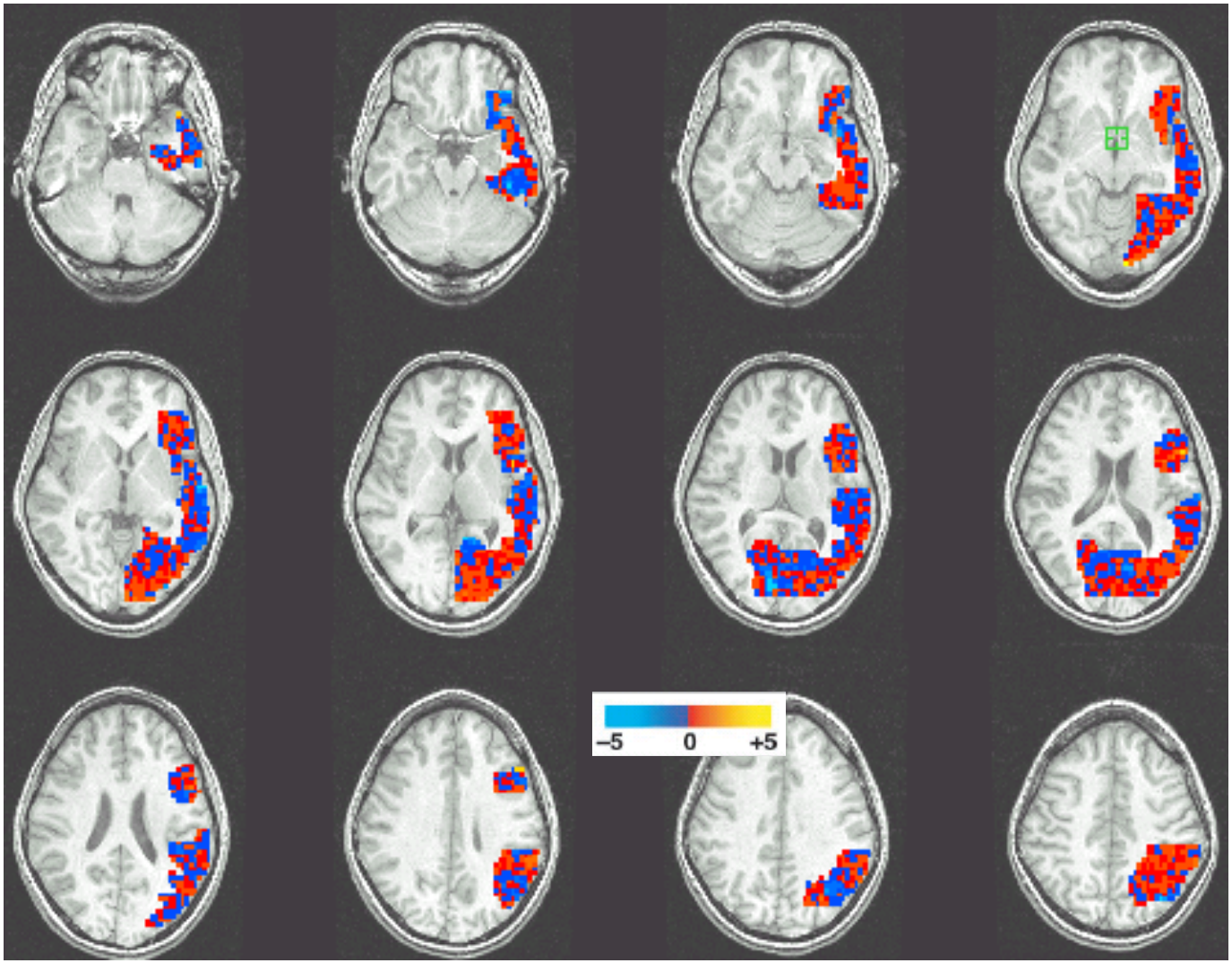
~2 images per sec.

15,000 voxels/image

non-invasive, safe

measures Blood Oxygen Level Dependent (BOLD) response

Gaussian Naïve Bayes: Learned $\mu_{voxel,word}$



[Mitchell et al.03]

15,000 voxels or features

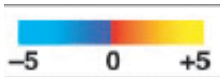
10 training examples or subjects per class (12 word categories)

Learned Naïve Bayes Models –
Means for P(BrainActivity | WordCategory)

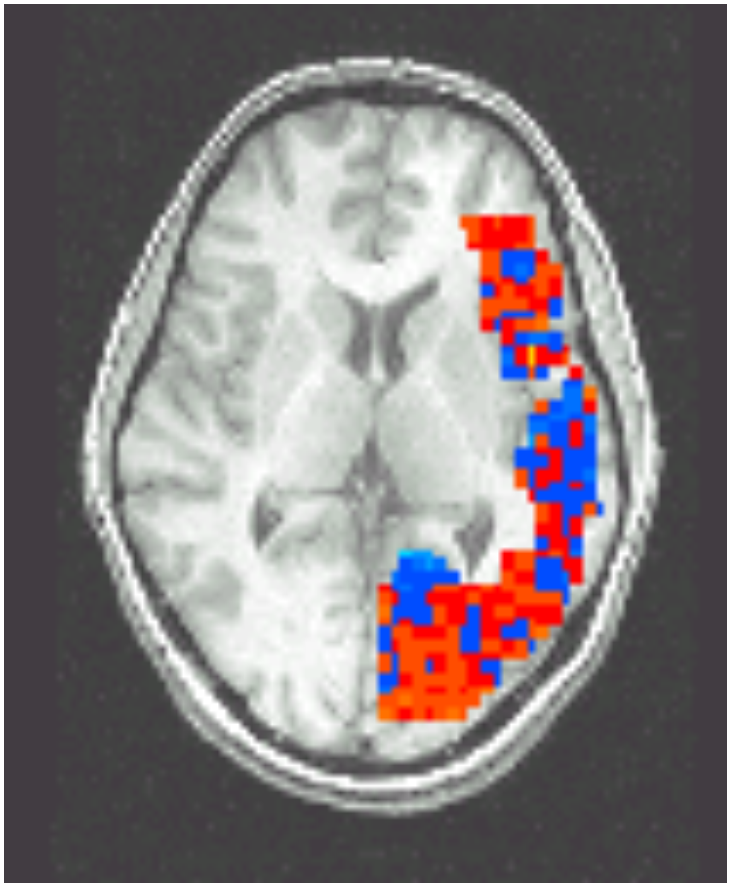
[Mitchell et al.03]

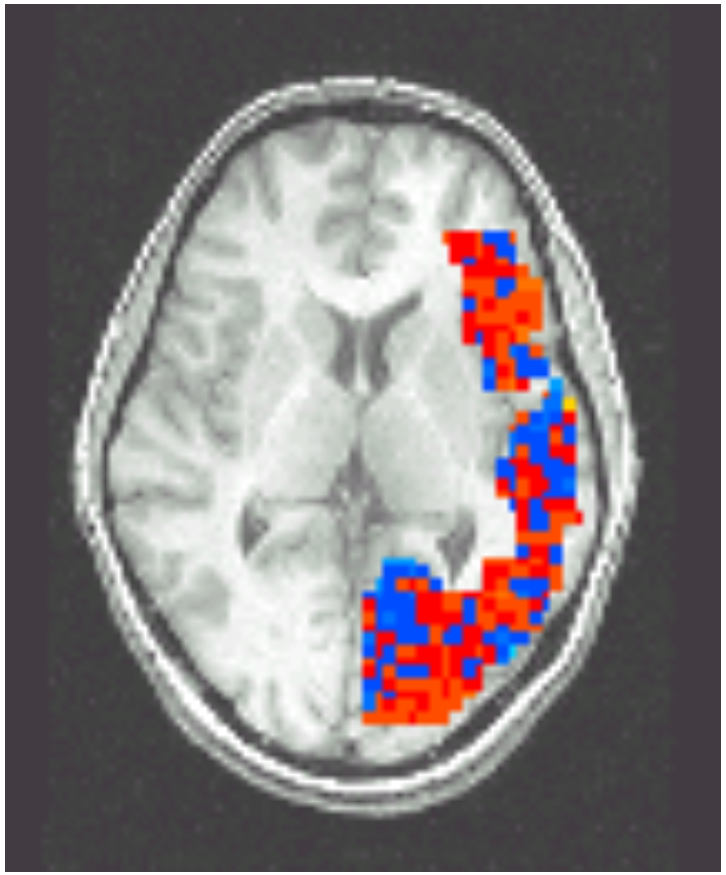
Pairwise classification accuracy: 85%

People words



Animal words





What you should know…

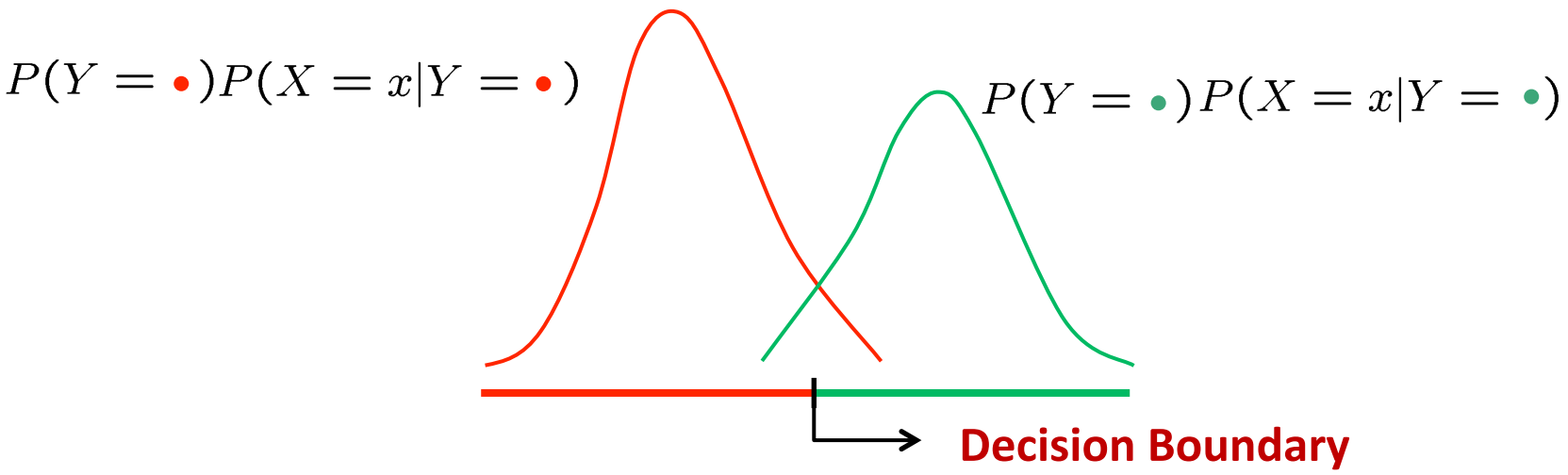
- Optimal decision using Bayes Classifier
- Naïve Bayes classifier
  - What's the assumption
  - Why we use it
  - How do we learn it (MLE, MAP)
  - Why is MAP estimation important

Decision Boundary of Bayes & Naïve Bayes Classifiers

- Binary classification with continuous features

  decision boundary is set of points x: P(Y=1|X=x) = P(Y=0|X=x)

If class conditional feature distribution P(X=x|Y=y) is 1-dim Gaussian $N(\mu,\sigma^2)$

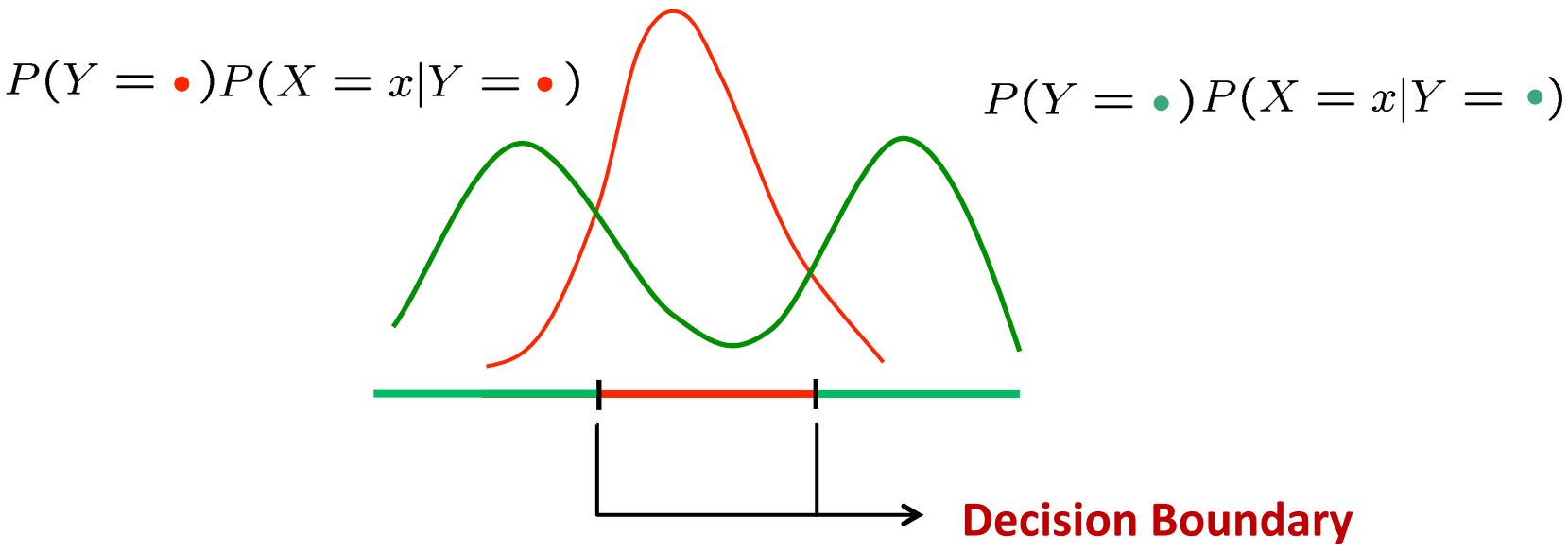


What other decision boundaries are possible in 1-dim be?

Decision Boundary of Bayes & Naïve Bayes Classifiers

- Binary classification with continuous features

  decision boundary is set of points x: P(Y=1|X=x) = P(Y=0|X=x)

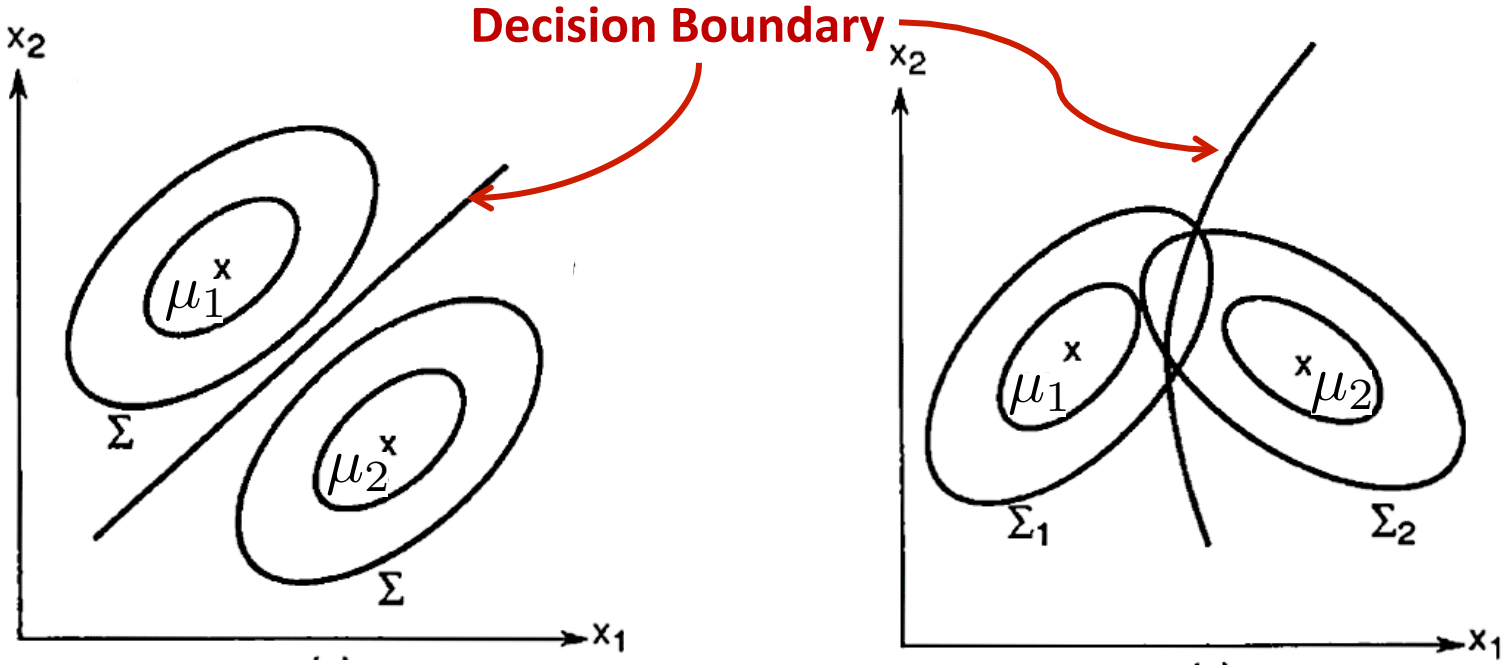


Decision Boundary of Bayes & Naïve Bayes Classifiers

- Binary classification with continuous features

  decision boundary is set of points x: P(Y=1|X=x) = P(Y=0|X=x)

If class conditional feature distribution P(X=x|Y=y) is 2-dim Gaussian $N(\mu_y,\Sigma_y)$



Decision Boundary of Bayes & Naïve Bayes Classifiers

- Binary classification with continuous features

  decision boundary is set of points x: P(Y=1|X=x) = P(Y=0|X=x)

If class conditional feature distribution P(X=x|Y=y) is 2-dim Gaussian $N(\mu_y,\Sigma_y)$

$$P(X = x|Y = y) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_y|}} \exp\left(-\frac{(x-\mu_y)\Sigma_y^{-1}(x-\mu_y)'}{2}\right)$$

$$\frac{P(Y=1|X=x)}{P(Y=0|X=x)} = \frac{P(X=x|Y=1)P(Y=1)}{P(X=x|Y=0)P(Y=0)}$$

$$= \sqrt{\frac{|\Sigma_0|}{|\Sigma_1|}} \exp\left(-\frac{(x-\mu_1)\Sigma_1^{-1}(x-\mu_1)'}{2} + \frac{(x-\mu_0)\Sigma_0^{-1}(x-\mu_0)'}{2}\right) \frac{\theta}{1-\theta}$$

Note: In general, this implies a quadratic equation.

But if $\Sigma_1 = \Sigma_0$, then quadratic part cancels out and equation is linear.