Stat 406: Algorithms for classification and prediction

Midterm review

Kevin Murphy

Mon 12 February, 2007[1]

[1] Slides last updated on February 10, 2007

Outline

- Some useful probability distributions
- Maximum likelihood estimation
- Bayesian estimation
- Empirical Bayes
- Bayesian classifiers

Probability distributions

Type	1D	MultiD
D	Binomial $Bin(x\|n,\theta)$	Multinomial $Mu(x\|n,\theta)$
D	Bernoulli $Ber(x\|\theta)$	Multinomial $Mu(x\|1,\theta)$
C	Gaussian $\mathcal{N}(x\|\mu,\sigma^2)$	Gaussian $\mathcal{N}(x\|\mu,\Sigma)$
C	Beta $Beta(x\|\alpha_1,\alpha_0)$	Dirichlet $Dir(x\|\alpha_1,\ldots,\alpha_K)$

Discrete probability distributions

Binomial $X \in \{0, 1, \ldots, N\}$, $\theta \in [0, 1]$

$$Bi(X|n, \theta) = \binom{n}{X} \theta^X (1 - \theta)^{n-X} \tag{1}$$

Bernoulli $X \in \{0, 1\}$, $\theta \in [0, 1]$

$$Be(X|\theta) = \theta^X (1 - \theta)^{1-X} = \theta^{I(X=1)} (1 - \theta)^{I(X=0)} \tag{2}$$

Multinomial $X_k \in \{0, 1, \ldots, N\}$, $\theta_k \in [0, 1]$, $\sum_{k=1}^K \theta_k = 1$

$$Mu(X|n, \theta) = \binom{n}{x_1 \ldots x_K} \prod_{j=1}^K \theta_j^{x_j} \tag{3}$$

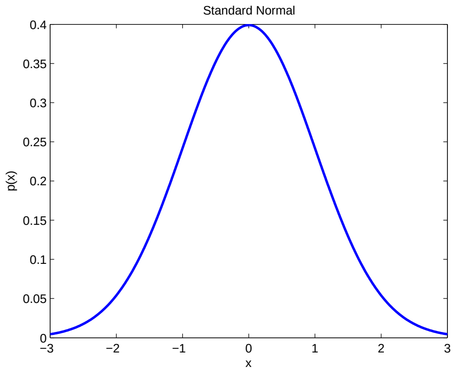
Multinomial $X \in \{1, \ldots, K\}$, $\theta_k \in [0, 1]$, $\sum_{k=1}^K \theta_k = 1$

$$Mu(X|1, \theta) = \prod_{j=1}^K \theta_j^{I(X=j)} \tag{4}$$

1D Gaussians

Univariate $x \in \mathbb{R}$, $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$

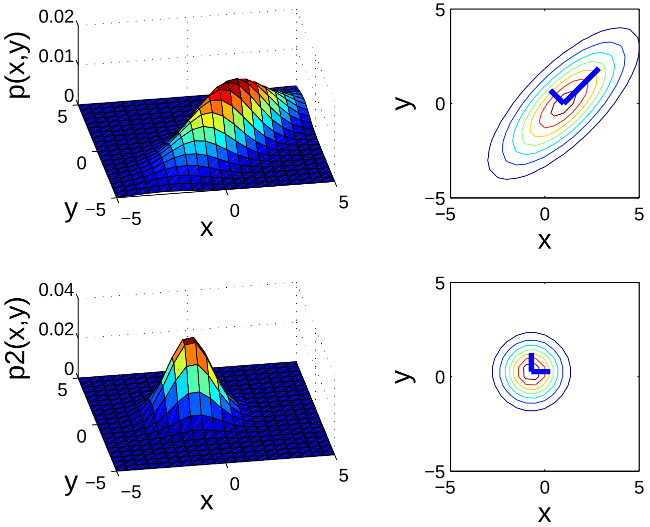
$$\mathcal{N}(x|\mu,\sigma) \overset{\text{def}}{=} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \tag{5}$$



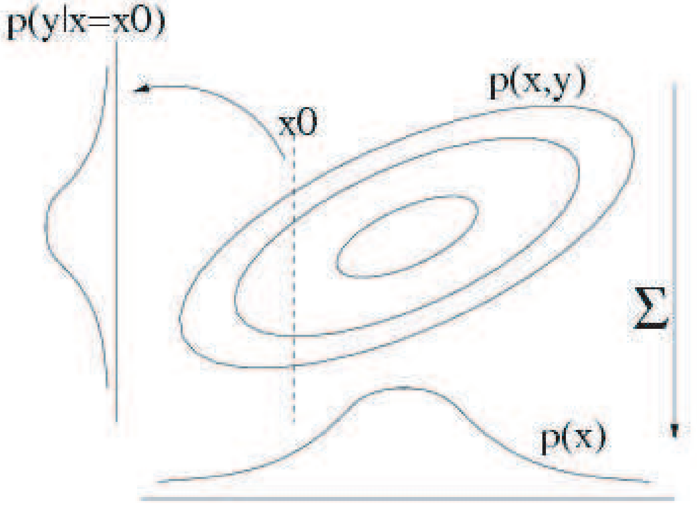
Multivariate Normal

MVN $x \in \mathbb{R}^d$, $\mu \in \mathbb{R}^d$, Σ psd

$$\mathcal{N}(\vec{x}|\vec{\mu},\Sigma) \stackrel{\text{def}}{=} \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp[-\tfrac{1}{2}(\vec{x}-\vec{\mu})^T\Sigma^{-1}(\vec{x}-\vec{\mu})] \qquad (6)$$



MVN: marginals and conditionals



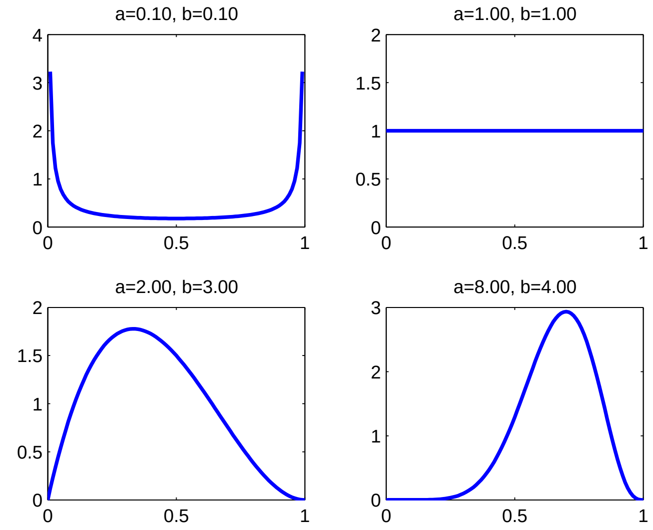
$$
\begin{aligned}
p(x_1, x_2) &= p(x_2)p(x_1|x_2) && (7)\\
&= \mathcal{N}(x_2|\mu_2, \Sigma_{22})\mathcal{N}(x_1|\mu_{1|2}, \Sigma_{1|2}) && (8)
\end{aligned}
$$

$$
\begin{aligned}
\mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) && (9)\\
\Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} && (10)
\end{aligned}
$$

Beta distribution

Beta $x \in [0,1]$, $\alpha_0, \alpha_1 \in \mathbb{R}^+$

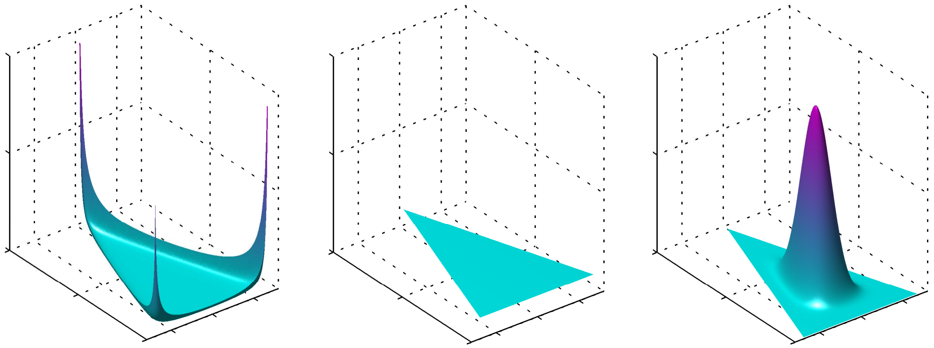
$$\text{Beta}(x|\alpha_1, \alpha_0) = \frac{1}{B(\alpha_1, \alpha_0)} x^{\alpha_1 - 1}(1-x)^{\alpha_0 - 1} \tag{11}$$



Dirichlet distribution

Dirichlet $x \in [0,1]^K$, $\sum_j x_j = 1$, $\alpha_j \in \mathbb{R}^+$

$$\mathcal{D}(x|\alpha) = \frac{1}{Z(\alpha)} \cdot x_1^{\alpha_1 - 1} \cdot x_2^{\alpha_0 - 1} \cdots x_K^{\alpha_K - 1} \tag{12}$$



Outline

- Some useful probability distributions $\surd$
- Maximum likelihood estimation
- Bayesian estimation
- Empirical Bayes
- Bayesian classifiers

Maximum likelihood estimation

$$\hat{\theta}^{mle} = \arg\max_{\theta} \log \prod_{i=1}^{n} p(x_i|\theta) = \arg\max_{\theta} \sum_{i=1}^{n} \log p(x_i|\theta) \quad (13)$$

- Most widely used point estimate (enjoys various theoretical properties, often gives intuitive answer)
- When maximizing likelihood with constraints, use Lagrange multipliers eg., for multinomial, we maximize

$$\tilde{\ell}(\theta) = \log \prod_{k} \theta_k^{N_k} + \lambda(1 - \sum_{k} \theta_k) \quad (14)$$

- Frequentist (classical) statistics represents uncertainty in estimates by deriving the sampling distribution $\hat{\theta}(X_{1:n})$ from the underyling distributions of the X_i; often this is asymptotically Gaussian. From this, we can derive confidence intervals.

Some MLEs

Bernoulli

$$\theta = \frac{\sum_i x_i}{n} \tag{15}$$

Multinomial

$$\theta_j = \frac{\sum_i I(x_i = j)}{n} \tag{16}$$

Gaussian

$$\mu = \frac{\sum_i x_i}{n} = \overline{x} \tag{17}$$

$$\sigma^2 = \frac{1}{n} \sum_i (x_i - \overline{x})^2 \tag{18}$$

$$\Sigma = \frac{1}{n} \sum_i (x_i - \overline{x})(x_i - \overline{x})^T = \frac{1}{n} X X^T - \overline{x}\overline{x}^T \tag{19}$$

Outline

- Some useful probability distributions $\surd$
- Maximum likelihood estimation $\surd$
- Bayesian estimation
- Empirical Bayes
- Bayesian classifiers

Bayesian parameter estimation

In Bayesian statistics, all forms of uncertainty are represented with probability distributions. We estimate parameters using Bayes rule

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta} \tag{20}$$

A natural conjugate prior $p(\theta)$ is a distribution of the same functional form as the likelihood $p(D|\theta)$. In this case, the posterior $p(\theta|D)$ is also the same form, and can be used as a prior for the next round of learning.

Beta-Bernoulli model

- Prior

$$p(\theta) = Be(\theta|\alpha_1, \alpha_0) = \frac{1}{B(\alpha_1, \alpha_0)}\theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1} \quad (21)$$

$$B(\alpha_1, \alpha_0) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_0)}{\Gamma(\alpha_1 + \alpha_0)} \quad (22)$$

- Likelihood

$$p(D|\theta) = \prod_{i=1}^{N} \theta^{N_1}(1-\theta)^{N_0} \quad (23)$$

- Posterior

$$p(\theta|D) = Be(\alpha_1 + N_1, \alpha_0 + N_0) = Be(\alpha_1', \alpha_0') \quad (24)$$

- Marginal likelihood:

$$p(D) = \frac{B(\alpha_0', \alpha_1')}{B(\alpha_1, \alpha_0)} \quad (25)$$

- Posterior predictive.

$$p(X = 1|D) = \frac{\alpha'_1}{\alpha'_1 + \alpha'_0} \tag{26}$$

DIRICHLET-MULTINOMIAL $(M_i = 1)$ MODEL

- Prior

$$p(\theta|\vec{\alpha}) \;=\; Dir(\theta|\vec{\alpha}) = \frac{1}{Z(\alpha)} \cdot \theta_1^{\alpha_1 - 1} \cdot \theta_2^{\alpha_2 - 1} \cdots \theta_K^{\alpha_K - 1} \qquad (27)$$

$$Z_{Dir}(\vec{\alpha}) \;=\; \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{K} \alpha_i)} \qquad (28)$$

- Likelihood

$$p(D|\vec{\theta}) \;=\; \prod_{j=1}^{K} \theta_j^{N_j} \qquad (29)$$

- Posterior

$$p(\theta|D,\vec{\alpha}) \;=\; Dir(\alpha_1 + N_1, \ldots, \alpha_K + N_K) = Dir(\alpha'_1, \ldots, \alpha'_K) \qquad (30)$$

- Marginal likelihood:

$$p(D) = \frac{Z_{Dir}(\vec{N}+\vec{\alpha})}{Z_{Dir}(\vec{\alpha})} \tag{31}$$

$$\tag{32}$$

- Posterior predictive

$$p(X=j|D) = \frac{\alpha_j + N_j}{N + \sum_k \alpha_k} \tag{33}$$

Normal-normal model

- Prior

$$p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0) = \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right) \quad (34)$$

- Likelihood

$$P(D|\mu, \sigma^2) = \prod_{i=1}^{N} \mathcal{N}(x_i|\mu, \sigma^2) \propto \mathcal{N}(\overline{x}|\mu, \sigma^2/N) \quad (35)$$

- Posterior

$$p(\mu|D, \sigma^2) = \mathcal{N}(\mu|\mu_N, \sigma_N^2) \quad (36)$$

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML} \quad (37)$$

$$\sigma_N^2 = \frac{\sigma^2\sigma_0^2}{N\sigma_0^2 + \sigma^2} \quad (38)$$

- Marginal likelihood

$$p(D) = \mathcal{N}(\overline{x}|\mu_0, \sigma_0^2 + \sigma^2/N) \tag{39}$$

- Posterior predictive

$$p(X|D, \sigma^2) = \mathcal{N}(X|\mu_N, \sigma_N^2 + \sigma^2) \tag{40}$$

Posterior mean

Posterior mean $E[\theta|D]$ is a common point estimate derived from the posterior. It is convex combination of prior mean and MLE. For Beta-Binomial:

$$E[\theta|D] = w\frac{\alpha_1}{\alpha_1+\alpha_0} + (1-w)\frac{N_1}{N_1+N_0} \tag{41}$$

where $w = (\alpha_1+\alpha_0)/(N_1+N_0+\alpha_1+\alpha_0)$.

For Normal-Normal:

$$E(\mu|D) = w\mu_0 + (1-w)\overline{x} \tag{42}$$

where $w = \frac{\lambda_0}{\lambda_0+N\lambda}$.

Posterior mode (MAP estimation)

Posterior mode is another common point estimate

$$\hat{\theta}^{MAP} = \arg\max p(\theta|D) = \arg\max \log p(D|\theta) + \log p(\theta) \tag{43}$$

This is equivalent to penalized (regularized) maximum likelhiood.

For a Gaussian, posterior mode = posterior mean.

For a Beta,

$$\hat{\theta}^{MAP} = \frac{N_1 + \alpha_1 - 1}{N_0 + N_1 + \alpha_1 + \alpha_0 - 2} \tag{44}$$

Outline

- Some useful probability distributions $\surd$
- Maximum likelihood estimation $\surd$
- Bayesian estimation $\surd$
- Empirical Bayes
- Bayesian classifiers

Empirical Bayes for a Gaussian mean

Suppose $X_i \sim \mathcal{N}(\theta_i, \sigma^2)$ and $\theta_i \sim \mathcal{N}(\mu, \tau^2)$. We will estimate the hyper parameters using ML-II:

$$\begin{aligned}
(\mu, \tau^2) &= \arg\max \prod_i p(X_i|\mu, \tau^2) && (45) \\
&= \arg\max \prod_i \int p(X_i|\theta_i) p(\theta_i|\mu, \tau^2) d\theta_i && (46) \\
&= \arg\max \prod_i \mathcal{N}(x_i|\mu, \tau^2 + \sigma^2) && (47)
\end{aligned}$$

We find

$$\begin{aligned}
\hat{\mu} &= \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x} && (48) \\
\hat{\tau}^2 &= s^2 - \sigma^2 && (49)
\end{aligned}$$

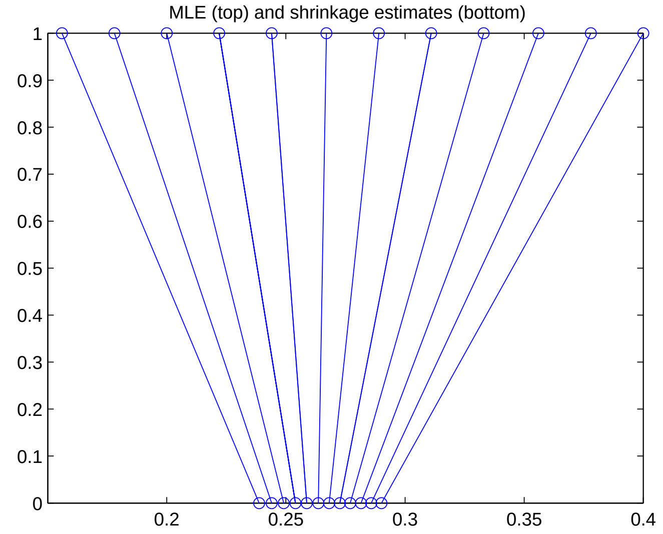
Shrinkage

Plugging in the estimated hyperparameters yields the following posterior mean:

$$\hat{\theta}_i = B\overline{x} + (1-B)x_i = \overline{x} + (1-B)(x_i - \overline{x}) \tag{50}$$

where the shrinkage factor is

$$\hat{B} = \frac{\sigma^2}{s^2} \tag{51}$$



Empirical Bayes for the Beta-Binomial model

Suppose $X_i \sim Bin(n_i, \theta_i)$ and $\theta_i \sim Beta(\alpha, \beta)$. We can estimate the hyper parameters using ML-II:

$$
\begin{aligned}
(\alpha, \beta) &= \arg\max \prod_i \int p(X_i|\theta_i) p(\theta_i|\alpha, \beta) d\theta_i && (52) \\
&= \arg\max \prod_i \frac{B(I(X_i = 1) + \alpha, I(X_i = 0) + \beta)}{B(\alpha, \beta)} && (53)
\end{aligned}
$$

or we can use method of moments to get

$$
\begin{aligned}
\hat{\alpha} &= \frac{m_1(m_2 - tm_1)}{m_1((t-1)m_1 + t) - tm_2} && (54) \\
\hat{\beta} &= \frac{(t - m_1)(m_2 - tm_1)}{m_1(4m_1 + t) - tm_2} && (55)
\end{aligned}
$$

where $m_1 = \frac{1}{n}\sum_{i=1}^n X_i$ and $m_2 = \frac{1}{n}\sum_{i=1}^n X_i^2$ and $t = n_i$.

Outline

- Some useful probability distributions $\surd$
- Maximum likelihood estimation $\surd$
- Bayesian estimation $\surd$
- Empirical Bayes $\surd$
- Bayesian classifiers

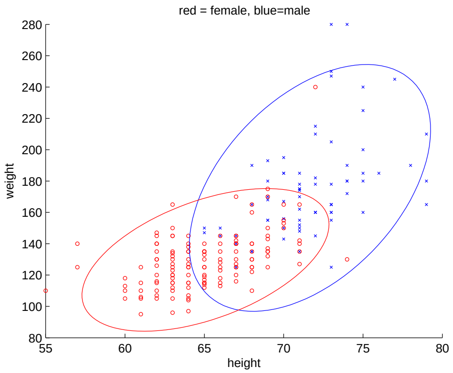
Bayesian (generative) classifiers

$$p(y|x) = \frac{p(x,y)}{p(x)} = \frac{p(x|y)p(y)}{\sum_{y'=1}^{C} p(x|y')p(y')} \tag{56}$$

$p(y)$ is Multinomial.

Class conditioanl densities $p(x|y)$ can be Gaussian or multinomial or other.

Gaussian class conditional densities



We usually use the plug-in rule

$$p(x|y=c) = \mathcal{N}(x|\hat{\mu}_c, \hat{\Sigma}_c) \tag{57}$$

where the MLEs are

$$\hat{\vec{\mu}}_c = \frac{1}{N_c} \sum_{i:y_i=c} \vec{x}_i \tag{58}$$

$$\hat{\Sigma}_c = \frac{1}{N_c} \sum_{i:y_i=c} (\vec{x}_i - \hat{\vec{\mu}}_c)(\vec{x}_i - \hat{\vec{\mu}}_c)' \tag{59}$$

But we could be more Bayesian and use

$$p(x|y=c) = \int\int \mathcal{N}(x|\mu_C, \Sigma_C)p(\mu_C)p(\Sigma_C)d\mu_C d\Sigma_C \tag{60}$$

This results in a multivariate T-distribution which has fatter tails.

Multinomial conditional densities

We make the naive Bayes assumption

$$p(x|y=c) = \prod_{j=1}^{d} p(x_j|y=c) \tag{61}$$

We usually use the plug-in rule

$$p(x_j|y=c) = p(x_j|y=c,\hat{\theta}_{jc}) = \prod_{k=1}^{K} \hat{\theta}_{jck}^{I(x_j=k)} \tag{62}$$

where the MLEs are

$$\hat{\theta}_{jck} = \frac{\sum_{i:y_i=c} I(x_{ij}=k)}{N_c} \tag{63}$$

But we could be more Bayesian and use

$$p(x_j|y=c) = \int Mu(x_j|\theta_{jc}) Dir(\theta_{jc}|\alpha_{jc}) d\theta_{jc} = \prod_{k=1}^{K} \overline{\theta}_{jck}^{I(x_j=k)} \tag{64}$$

where we plug in the posterior mean parameters

$$\overline{\theta}_{jck} = \frac{\sum_{i:y_i=c} I(x_{ij}=k) + \alpha_{jck}}{N_c + \sum_{k'} \alpha_{jck'}} \tag{65}$$

Class posterior for Gaussians

If the class-conditional densities are Gaussian, then the posterior becomes

$$p(Y = j|\vec{x}) = \frac{p(\vec{x}|Y = j)p(Y = j)}{\sum_{k=1}^{C} p(\vec{x}|Y = k)p(Y = k)} \tag{66}$$

$$= \frac{\pi_j \exp\left[-\frac{1}{2}(\vec{x} - \mu_j)^T \Sigma_j^{-1}(\vec{x} - \mu_j)\right]}{\sum_k \pi_k \exp\left[-\frac{1}{2}(\vec{x} - \mu_k)^T \Sigma_k^{-1}(\vec{x} - \mu_k)\right]} \tag{67}$$

2 class case gives rise to logistic regression

$$p(Y = 1|\vec{x}) = \sigma(\beta^T \vec{x} + \gamma) \tag{68}$$

where

$$\beta \overset{\text{def}}{=} \Sigma^{-1}(\mu_1 - \mu_0) \tag{69}$$

$$\gamma \overset{\text{def}}{=} -\tfrac{1}{2}(\mu_1 - \mu_0)^T \Sigma^{-1}(\mu_1 + \mu_0) + \log \frac{\pi_1}{\pi_0} \tag{70}$$

$$\sigma(z) \overset{\text{def}}{=} \frac{1}{1 + e^{-z}} = \frac{e^z}{e^z + 1} \tag{71}$$

K CLASS CASE GIVES RISE TO SOFTMAX REGRESSION

$$p(Y=j|\vec{x}) = \frac{\pi_j \exp\left[-\frac{1}{2}(\vec{x}-\mu_j)^T\Sigma_j^{-1}(\vec{x}-\mu_j)\right]}{\sum_k \pi_k \exp\left[-\frac{1}{2}(\vec{x}-\mu_k)^T\Sigma_k^{-1}(\vec{x}-\mu_k)\right]} \tag{72}$$

$$= \frac{\exp\left[\mu_j^T\Sigma^{-1}x - \frac{1}{2}\mu_j^T\Sigma^{-1}\mu_j + \log\pi_j\right]}{\sum_k \exp\left[\mu_k^T\Sigma^{-1}\vec{x} - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log\pi_k\right]} \tag{73}$$

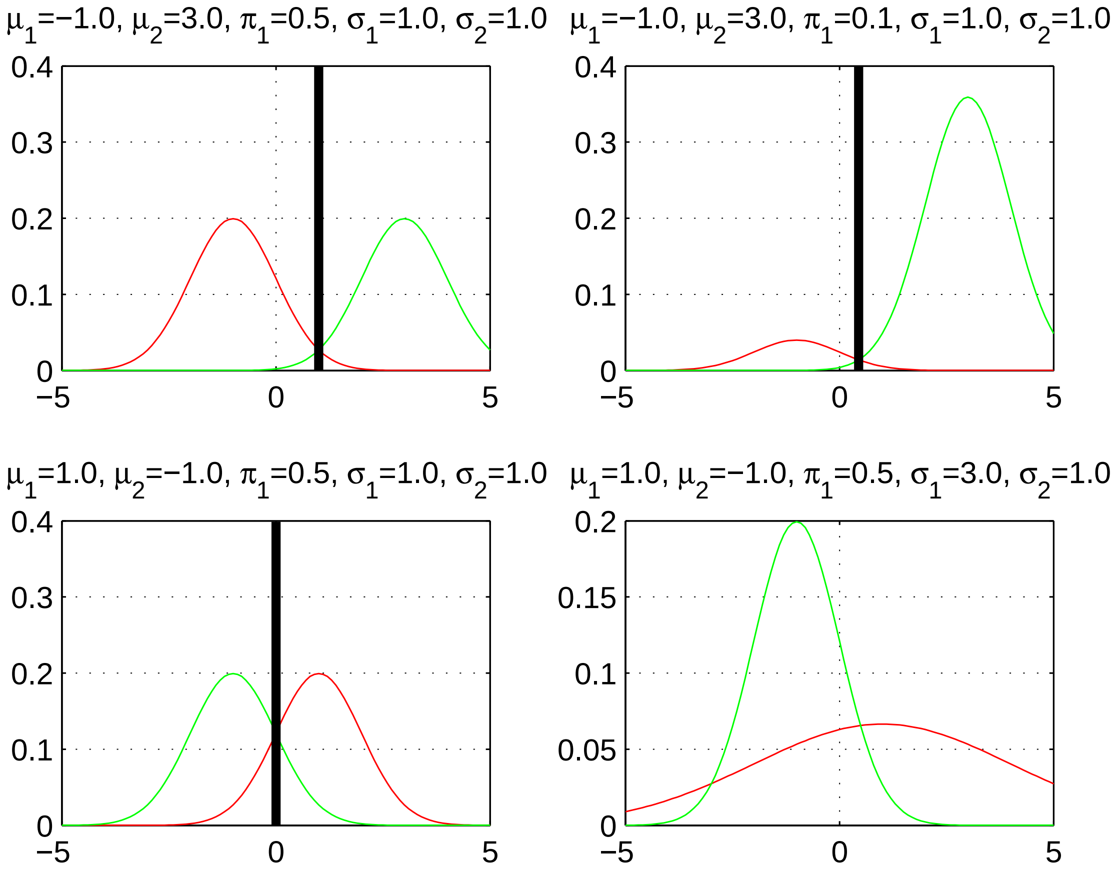
If we define

$$\theta_j \stackrel{\text{def}}{=} \begin{pmatrix} -\mu_j^T\Sigma^{-1}\mu_j + \log\pi_j \\ \Sigma^{-1}\mu_j \end{pmatrix} = \begin{pmatrix} \gamma_j \\ \beta_j \end{pmatrix} \tag{74}$$

then we find

$$p(Y=j|\vec{x}) = \frac{e^{\theta_j^T\vec{x}}}{\sum_k e^{\theta_k^T\vec{x}}} = \frac{e^{\beta_j^T\vec{x}+\gamma_j}}{\sum_k e^{\beta_k^T\vec{x}+\gamma_k}} \tag{75}$$

Decision boundary in 1D



μ_1=−1.0, μ_2=3.0, π_1=0.5, σ_1=1.0, σ_2=1.0

μ_1=−1.0, μ_2=3.0, π_1=0.1, σ_1=1.0, σ_2=1.0

μ_1=1.0, μ_2=−1.0, π_1=0.5, σ_1=1.0, σ_2=1.0

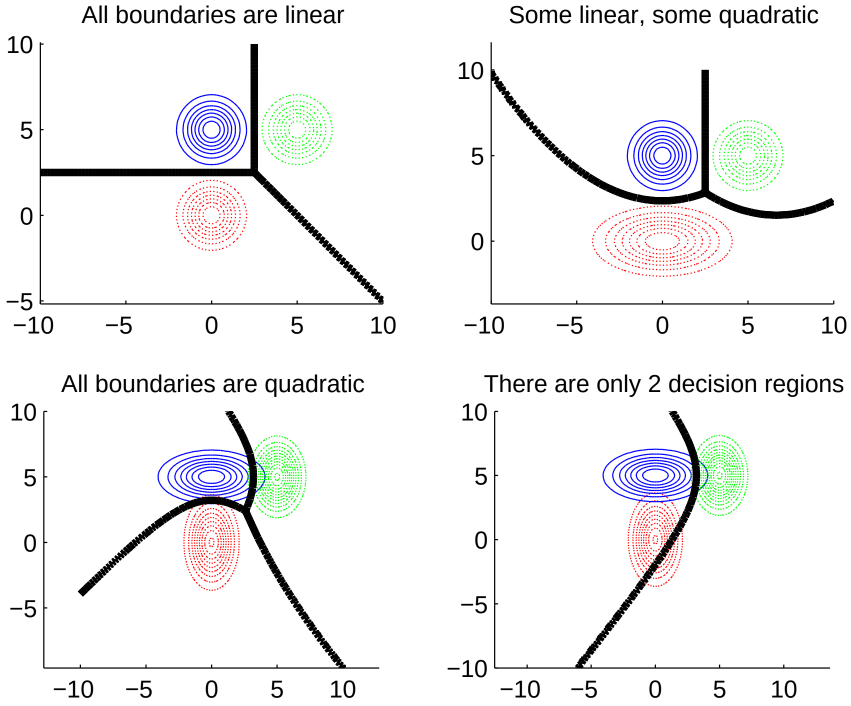
μ_1=1.0, μ_2=−1.0, π_1=0.5, σ_1=3.0, σ_2=1.0

Decision boundaries in 2D

The discriminant becomes

$$\begin{aligned} g_j(\vec{x}) &= \log p(\vec{x}|Y = j) + \log P(Y = j) \qquad (76) \\ &= -\tfrac{1}{2}(\vec{x} - \mu_j)^T \Sigma_j^{-1} (\vec{x} - \mu_j) - \frac{p}{2}\log(2\pi) - \tfrac{1}{2}\log|\Sigma_j| + \log p(Y = j) \qquad (77) \end{aligned}$$

If $\Sigma_c = \Sigma$ is shared, decision boundaries are linear, otherwise quadratic.


All boundaries are linear
Some linear, some quadratic
All boundaries are quadratic
There are only 2 decision regions