Discussion on Logistic Regression and Naïve Bayes

Jingrui He

09/27/2007

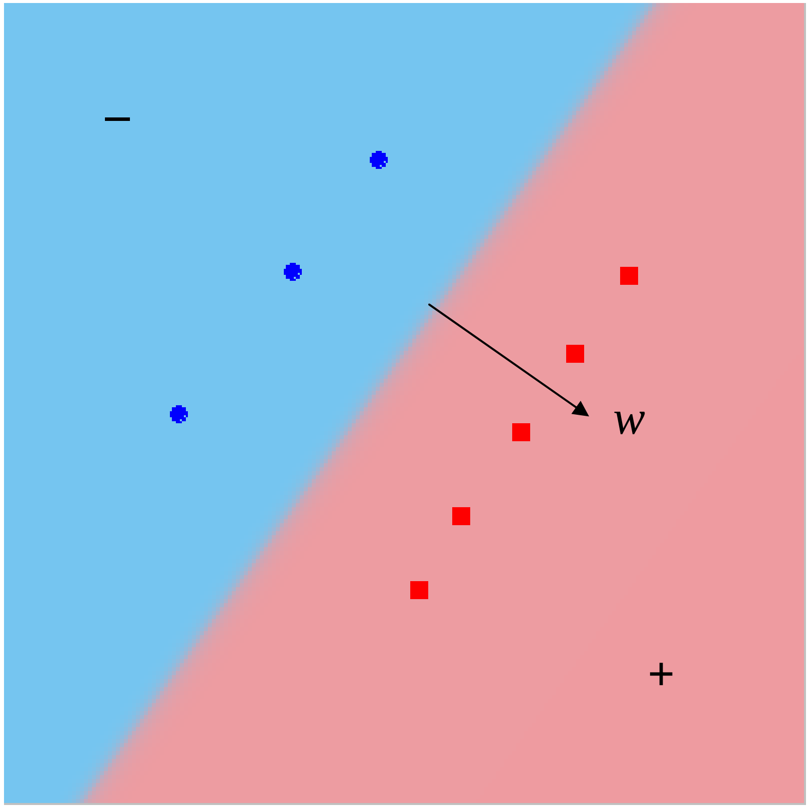
Review of Logistic Regression

- Discriminative classifier
- Function form for $P(Y|X)$
  - $$P(Y=1|X,w) = \frac{\exp\left(w_0 + \sum_i w_i X_i\right)}{1 + \exp\left(w_0 + \sum_i w_i X_i\right)}$$
- Can NOT obtain a sample of the data, because $P(X)$ is not available

Parameter Estimation

- Gradient ascent
  - $w_0^{t+1} \leftarrow w_0^t + \eta \sum_j \left[Y^j - \hat{P}\left(Y^j = 1 \middle| X^j, w^t\right) \right]$
  - $w_i^{t+1} \leftarrow w_i^t + \eta \sum_j X_i^j \left[Y^j - \hat{P}\left(Y^j = 1 \middle| X^j, w^t\right) \right]$
- Upon convergence
  - $\frac{\partial l(w)}{\partial w_0} = \sum_j \left[Y^j - P\left(Y^j = 1 \middle| X^j, w\right) \right] = 0$
  - $\frac{\partial l(w)}{\partial w_i} = \sum_j X_i^j \left[Y^j - P\left(Y^j = 1 \middle| X^j, w\right) \right] = 0$

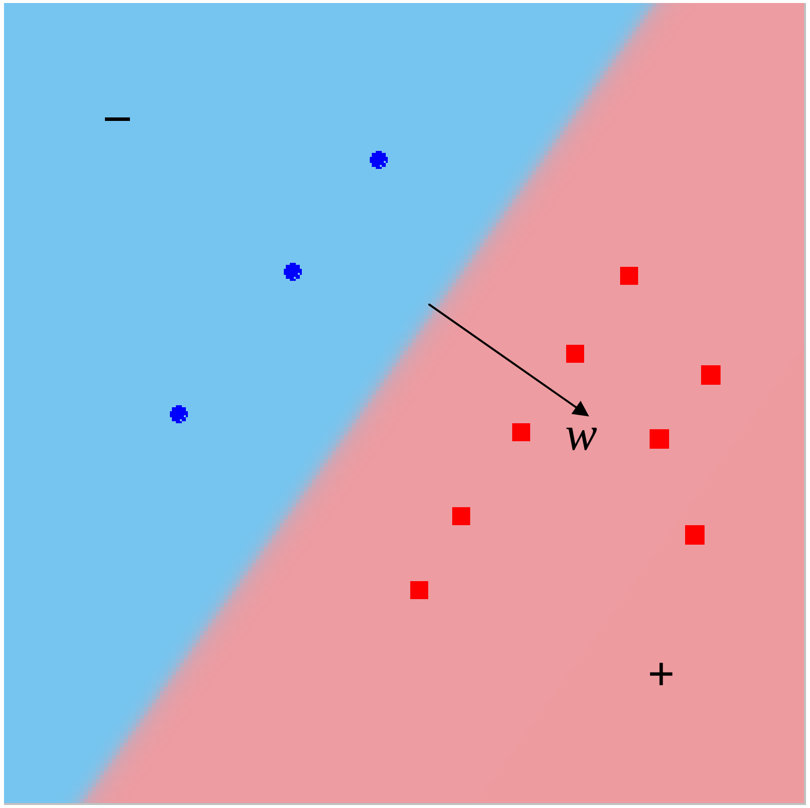
Linear Separable



- What's the value of *w*?
  - ***INFINITY!***
- Why?
  - Maximum likelihood

$$P(Y=1|X,w) = \frac{\exp\left(w_0 + \sum_i w_i X_i\right)}{1+\exp\left(w_0 + \sum_i w_i X_i\right)}$$

More Training Examples

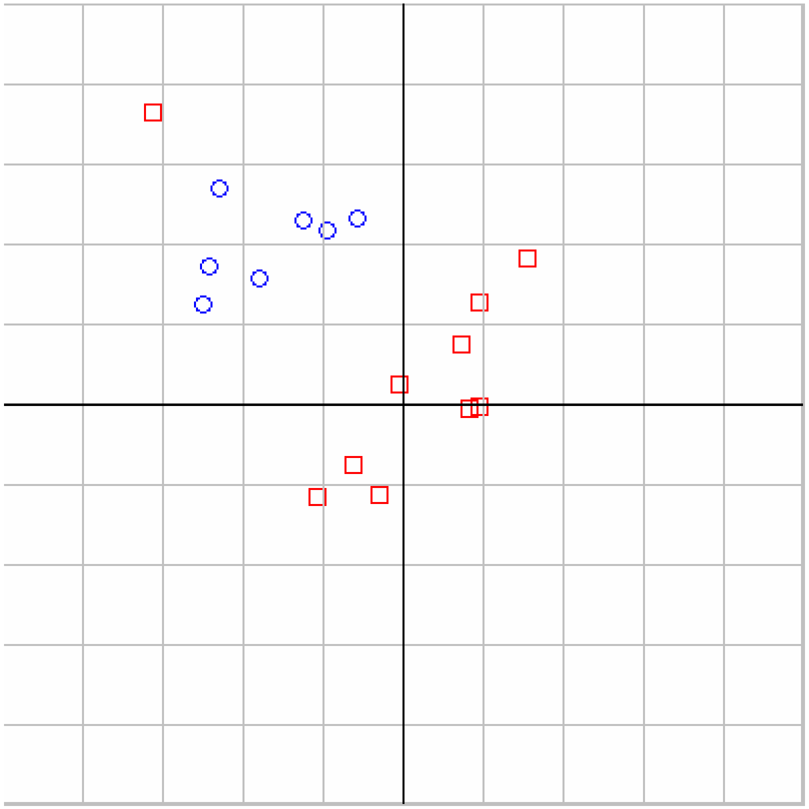


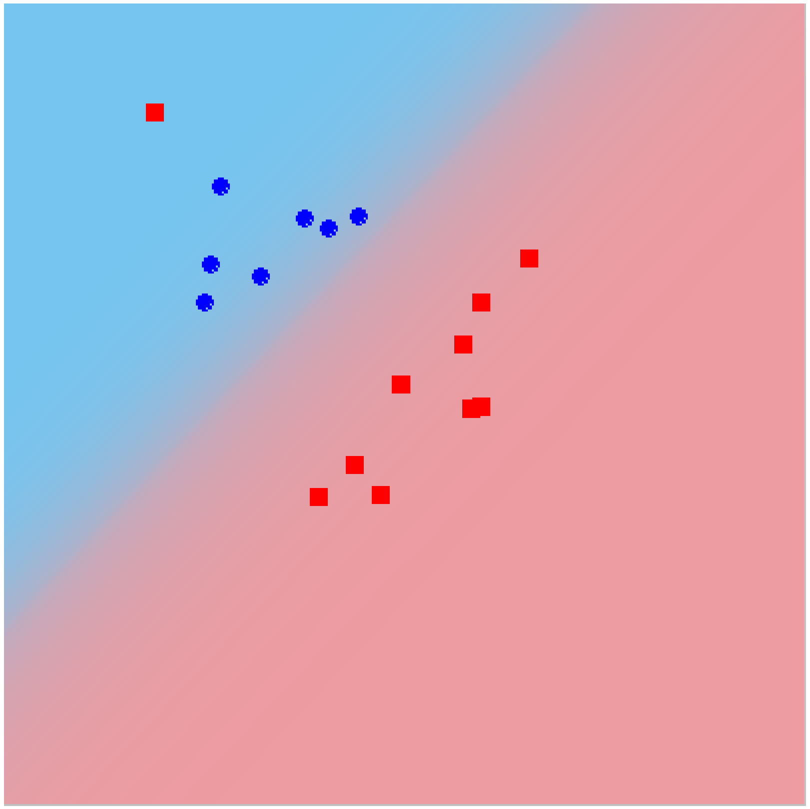
- No change in *w*
- Why?

$$w_0^{t+1} \leftarrow w_0^t + \eta \sum_j \left[Y^j - \hat{P}\left(Y^j = 1 \middle| X^j, w^t\right) \right]$$

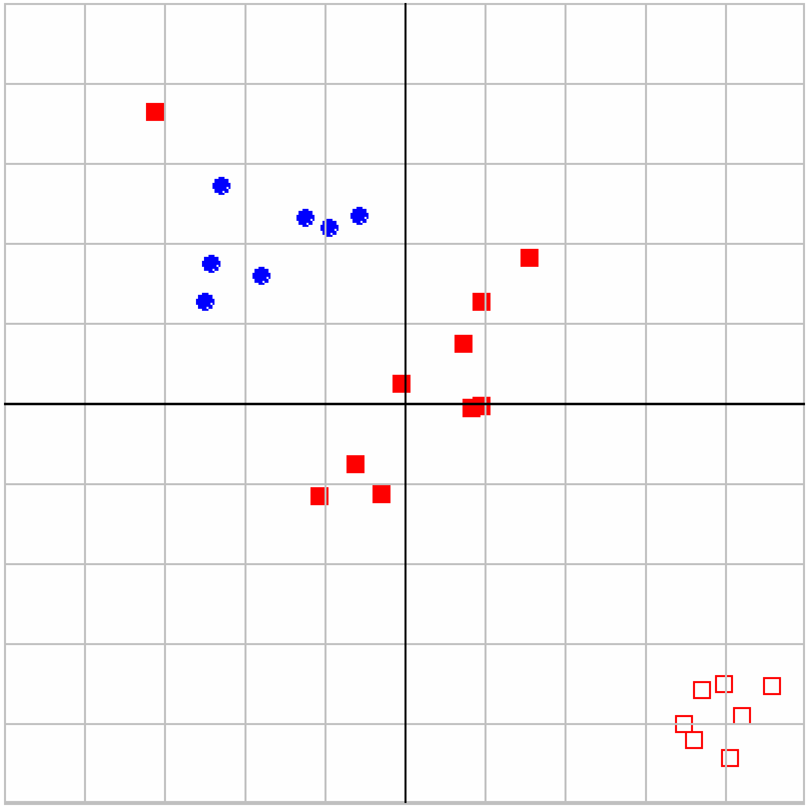
$$w_i^{t+1} \leftarrow w_i^t + \eta \sum_j X_i^j \left[Y^j - \hat{P}\left(Y^j = 1 \middle| X^j, w^t\right) \right]$$

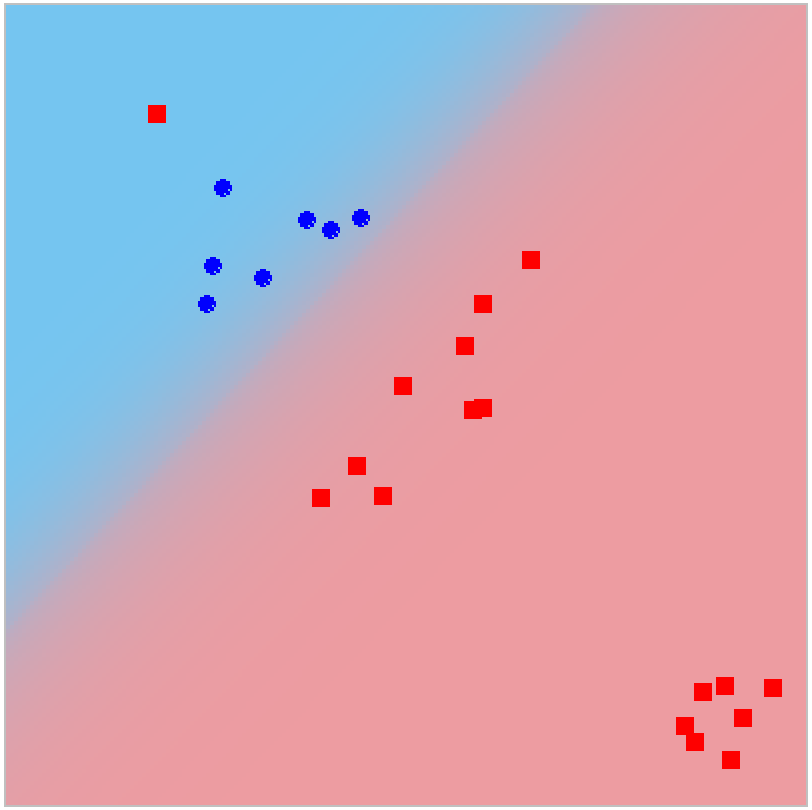
Non-Linear Separable



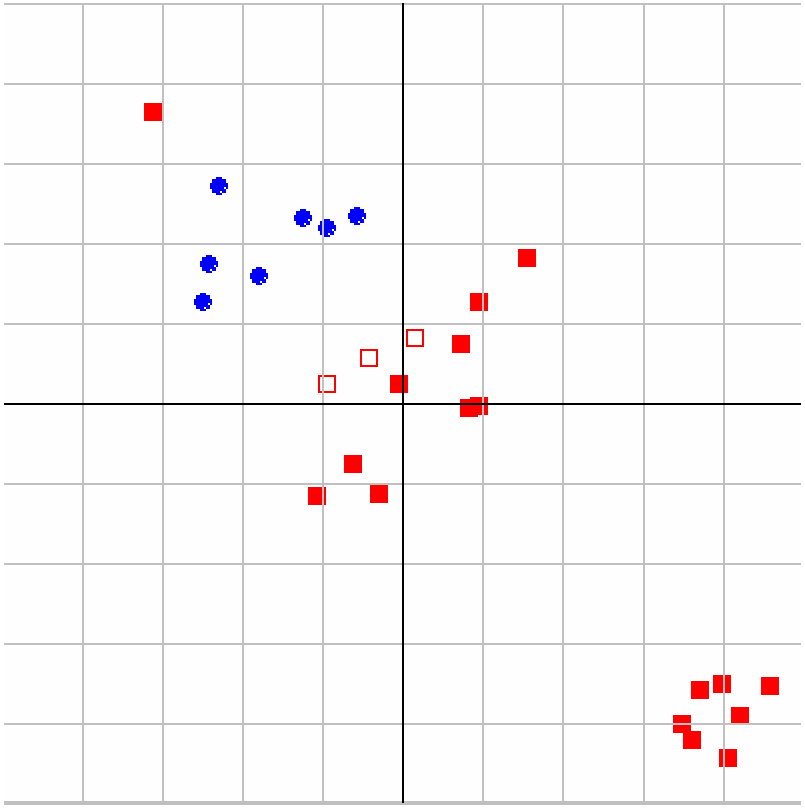


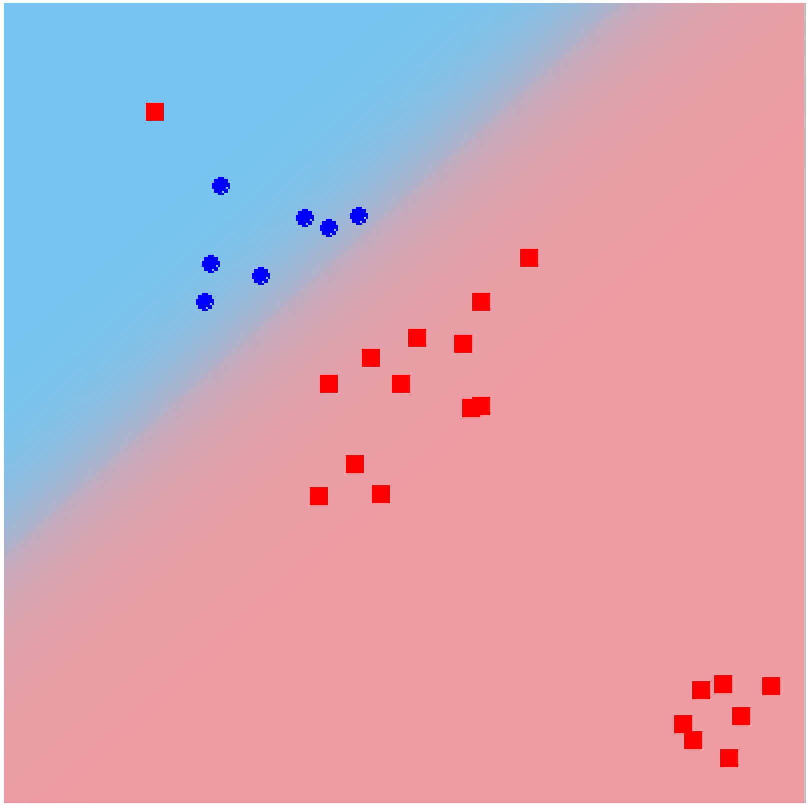
More Training Examples





Still More Training Examples





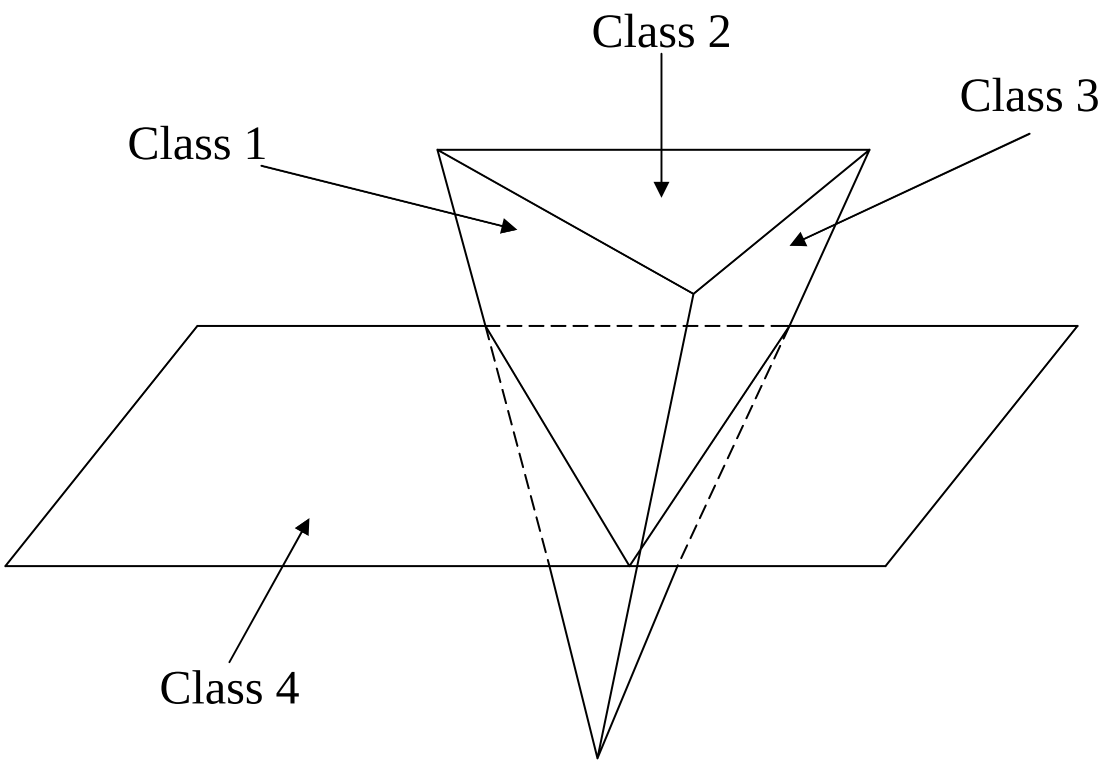
Why?

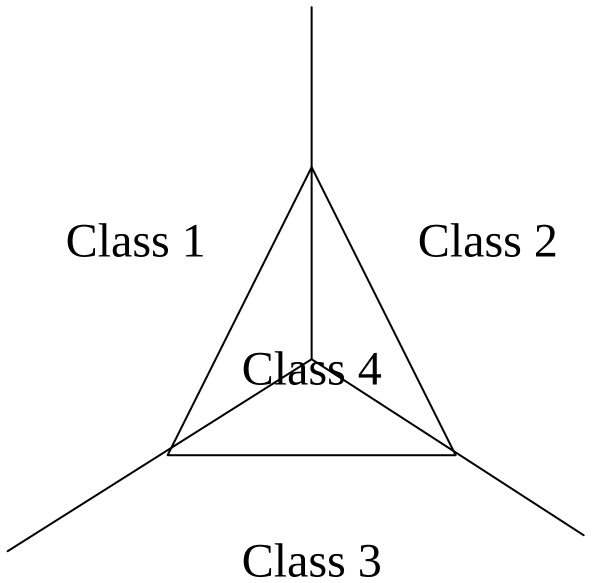
- Originally, upon convergence
  - $$\frac{\partial l(w)}{\partial w_0} = \sum_j \left[Y^j - P\left(Y^j = 1 \middle| X^j, w\right) \right] = 0$$
- With 3 more points
  - $$\frac{\partial l(w)}{\partial w_0} > 0$$
- To let the derivative be 0 again
  - Increase $P\left(Y^j = 1 \middle| X^j, w\right)$

Multiple Classes

- *R-1* sets of weights
  - $P\left(Y=j \middle| X, w_j\right) \propto \exp\left(w_{j0}+\sum_i w_{ji} X_i\right), \quad j=1,\ldots,R-1$
  - $P\left(Y=R \middle| X, w_j\right) = \dfrac{1}{1+\sum_{j=1}^{R-1} \exp\left(w_{j0}+\sum_i w_{ji} X_i\right)}$
- Classification
  - Comparing $\exp\left(w_{j0}+\sum_i w_{ji} X_i\right)$ and 1
  - Comparing $w_{j0}+\sum_i w_{ji} X_i$ and 0

4 Classes in 2d Space





LR vs. NB

- Loss functions
  - LR: maximum conditional data likelihood

    $$\sum_j \ln\left(P\left(Y^j \middle| X^j, w\right)\right)$$

  - NB: maximum data likelihood

    $$\sum_j \ln\left(P\left(X^j, Y^j \middle| w\right)\right)$$

- Different solutions!

LR vs. NB

- In NB, assume class independent variance

$$P\left(Y=1\middle|X,w\right)=\frac{1}{1+\exp\left(w_0+\sum_i w_i x_i\right)}$$

$$\ln\frac{1-\theta}{\theta}+\sum_i \frac{\mu_{i1}^2-\mu_{i0}^2}{2\sigma_i^2}$$

$$\frac{\mu_{i0}-\mu_{i1}}{\sigma_i^2}$$

LR vs. NB