# CS578 Statistical Machine Learning Lecture 12

Jean Honorio
Purdue University

*(based on slides by Tommi Jaakkola, MIT CSAIL)*

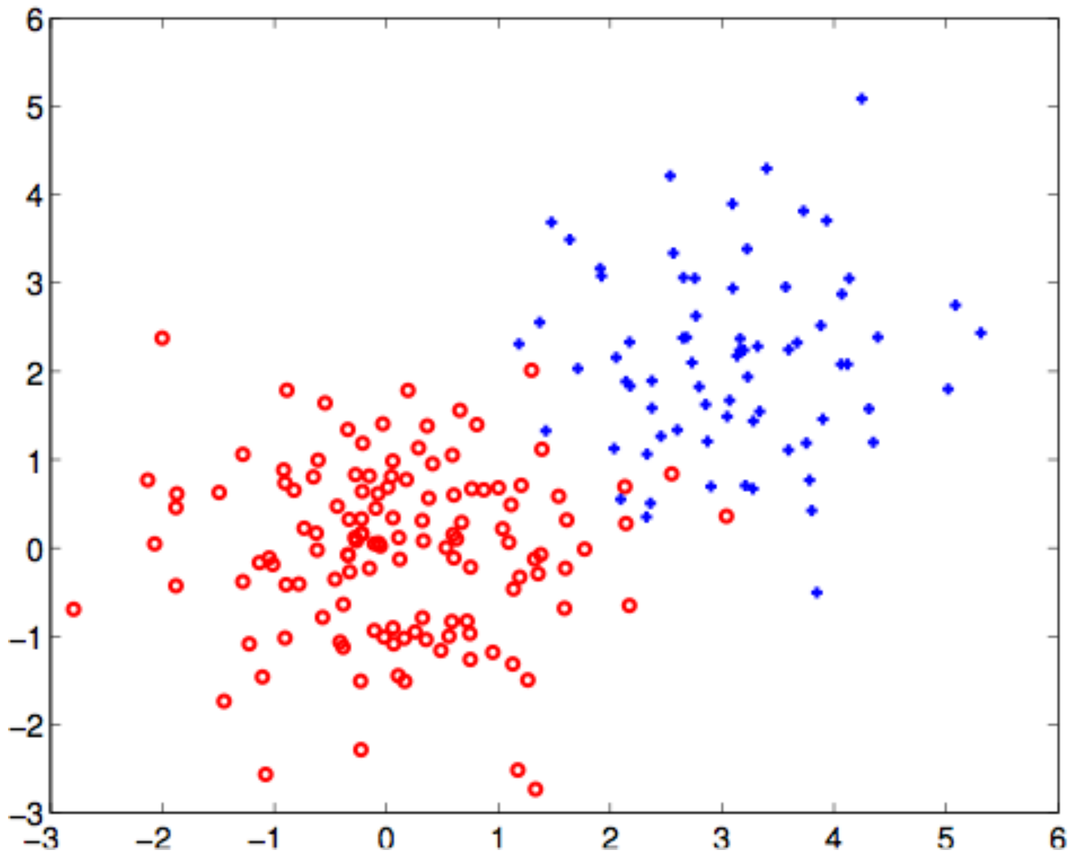
# Today's topics

- Generative probabilistic modeling
  - conditional Gaussian classifiers
- Maximum likelihood estimation
- Classification and decision boundary

# Training/test data generation

- We assume that the training (and test) examples are drawn as samples (generated) from some unknown distribution

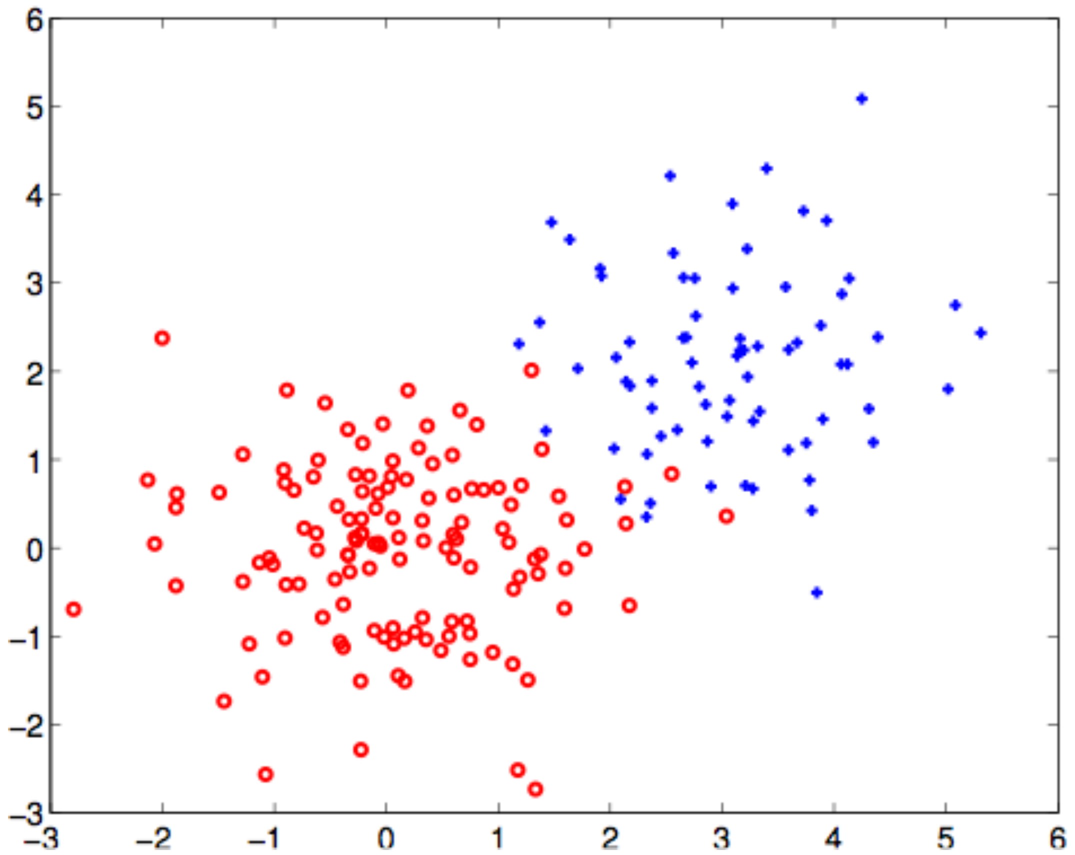
$$(\underline{x}, y) \sim P(\underline{x}, y)$$



# Training/test data generation

- We assume that the training (and test) examples are drawn as samples (generated) from some unknown distribution

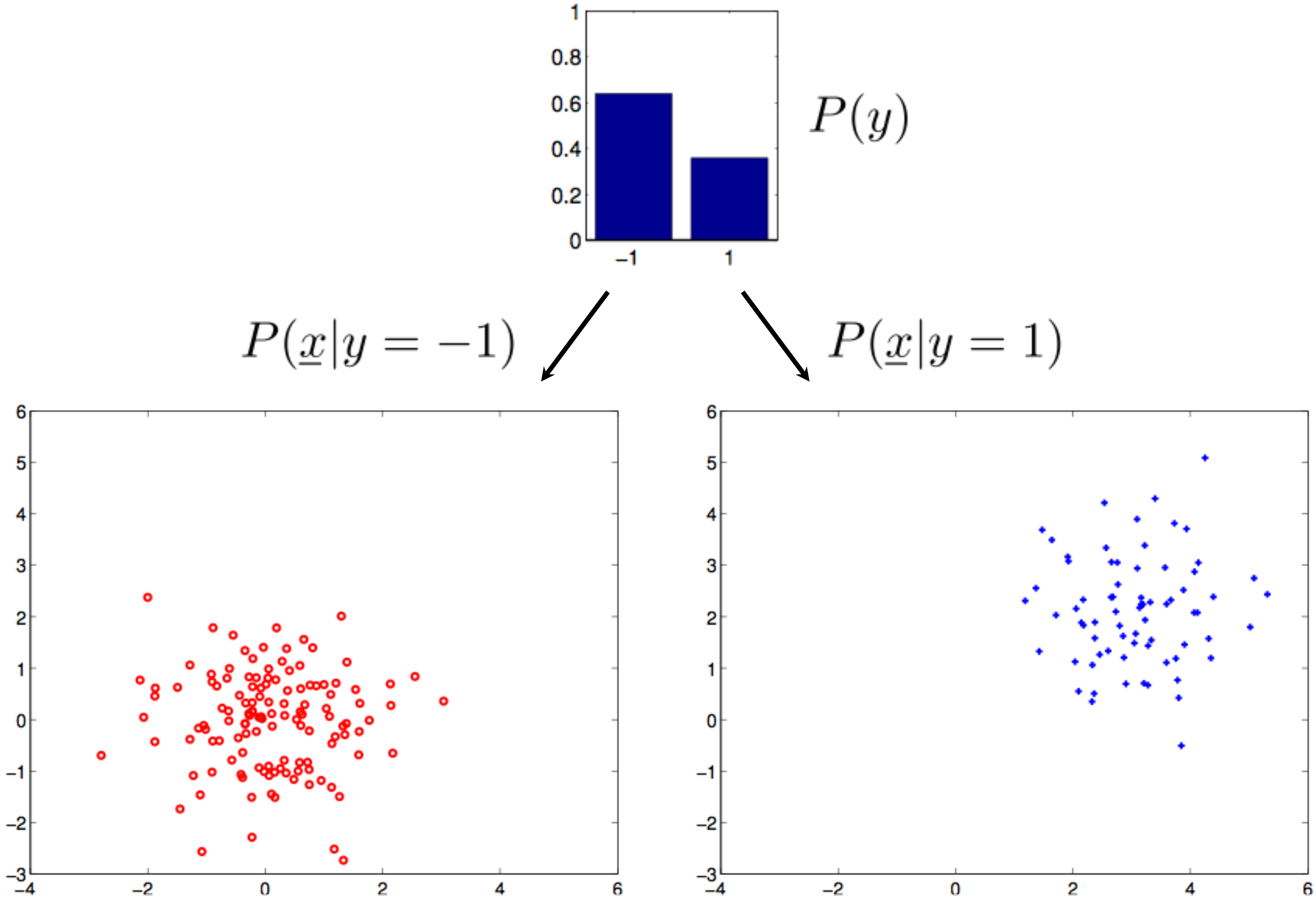
$$(\underline{x}, y) \sim P(\underline{x}, y) = P(\underline{x})P(y|\underline{x}) = P(\underline{x}|y)P(y)$$



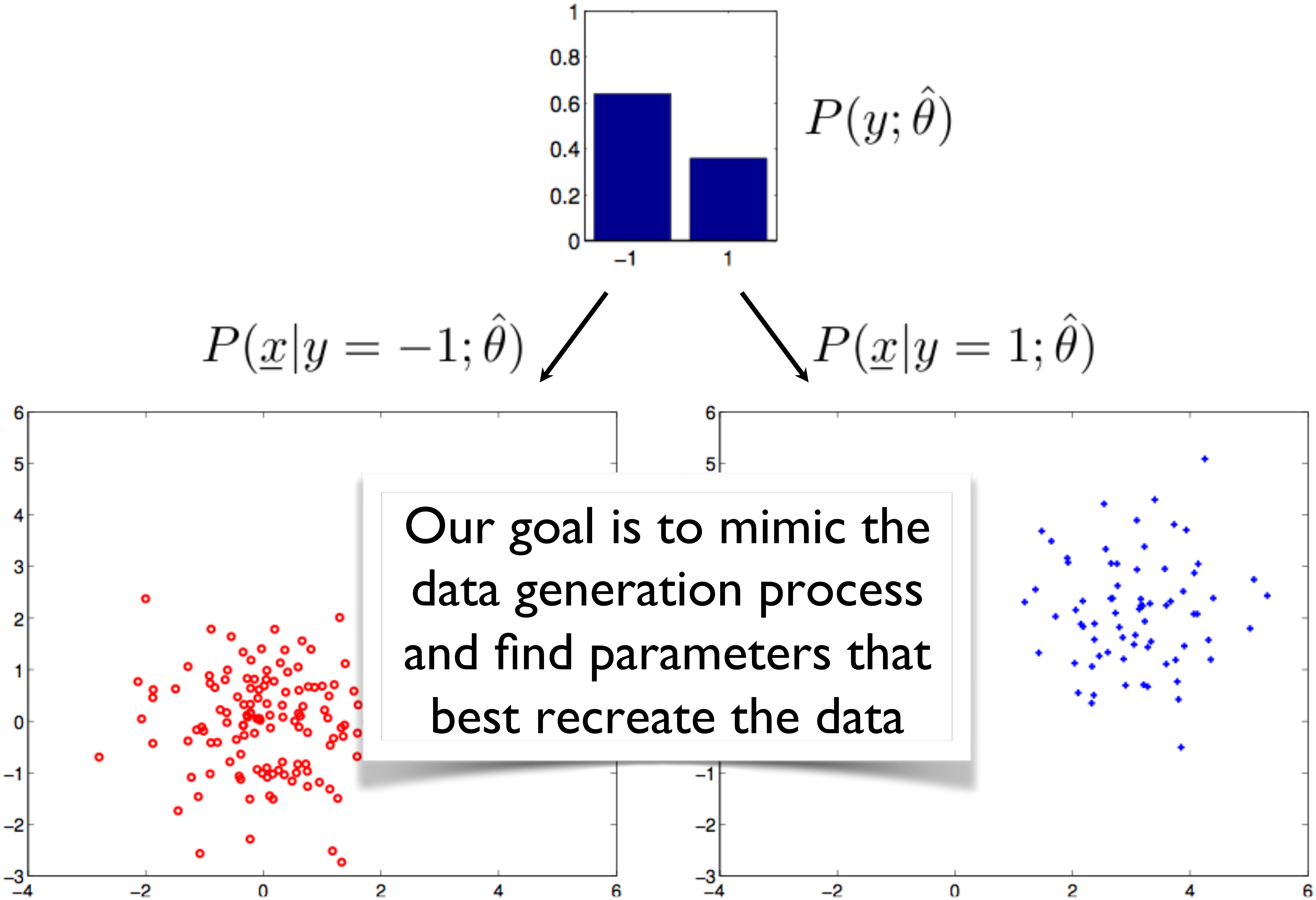

- We can always think of these samples $(\underline{x}, y)$ as having been generated in two steps: first $y$, then $\underline{x}$ given $y$

$$y \sim P(y), \quad \underline{x} \sim P(\underline{x}|y)$$

# Training/test data generation



# Generative modeling




Our goal is to mimic the data generation process and find parameters that best recreate the data

# The two approaches

- There are two broad approaches to classification problems:

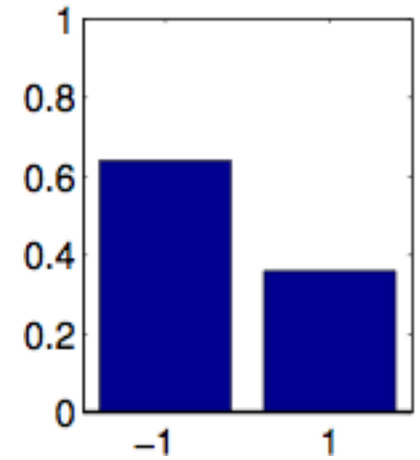
**Discriminative** (so far)

- model $=$ a set of classifiers
- choose $f$ that classifies training examples well
- label new inputs $\underline{x}$ based on $y = f(\underline{x})$

**Generative** (preview)

- model $=$ a set of distributions $P(\underline{x}, y; \theta)$ for any $\theta$
- choose $P(\underline{x}, y; \hat{\theta})$ such that training examples are likely samples from this distribution
- label new inputs $\underline{x}$ as $y$ where $y$ maximizes $P(\underline{x}, y; \hat{\theta})$

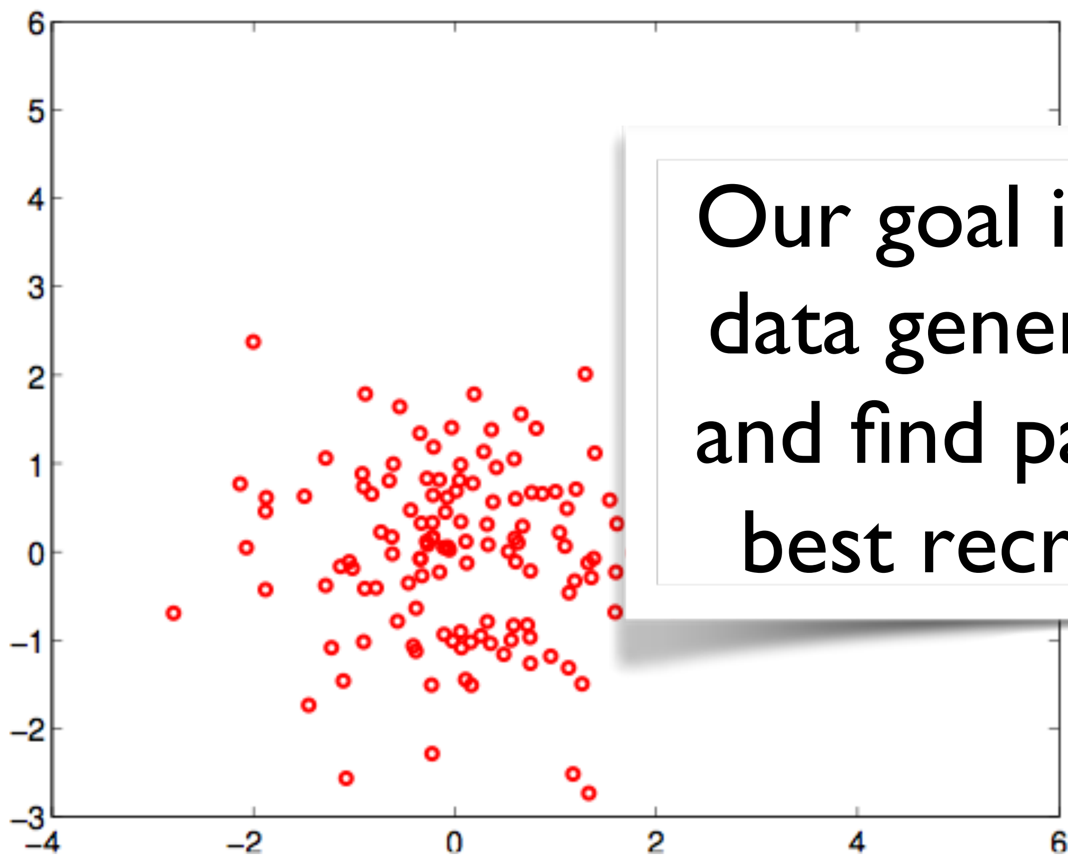
# Generative modeling

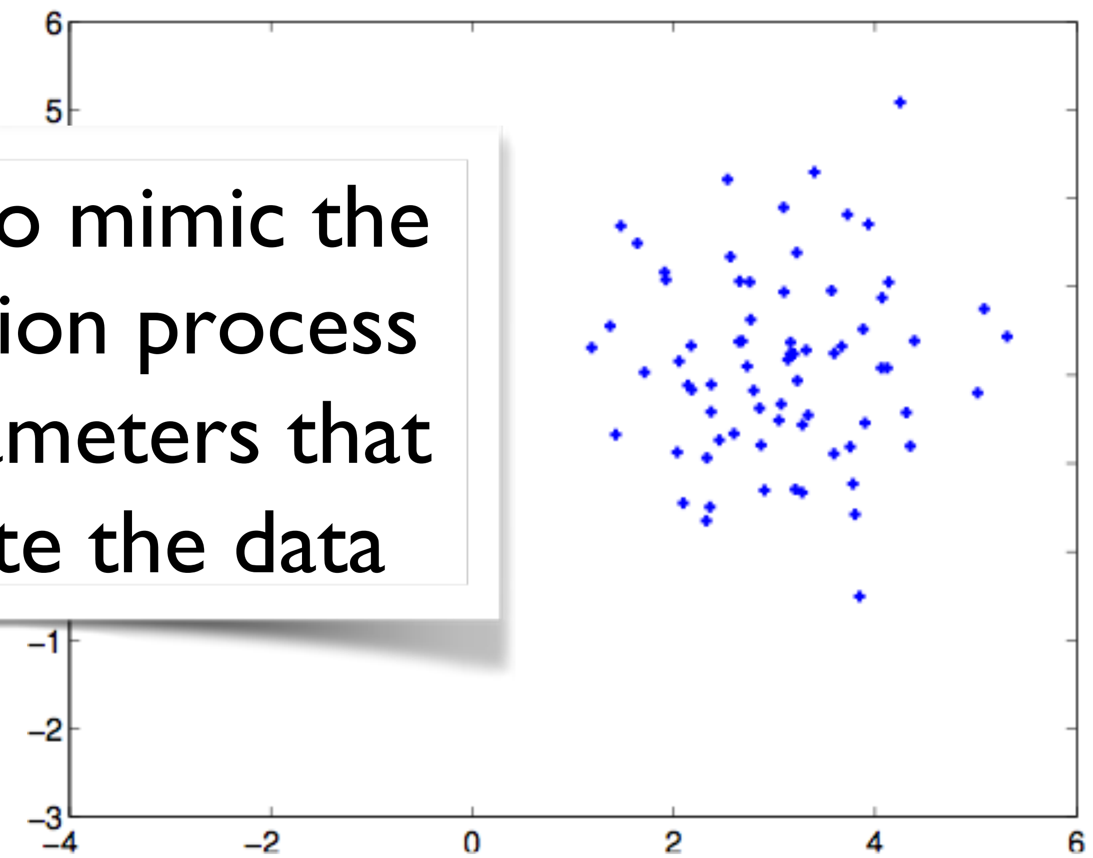


$P(y; \hat{\theta})$

$P(\underline{x}|y = -1; \hat{\theta})$

$P(\underline{x}|y = 1; \hat{\theta})$





Our goal is to mimic the data generation process and find parameters that best recreate the data

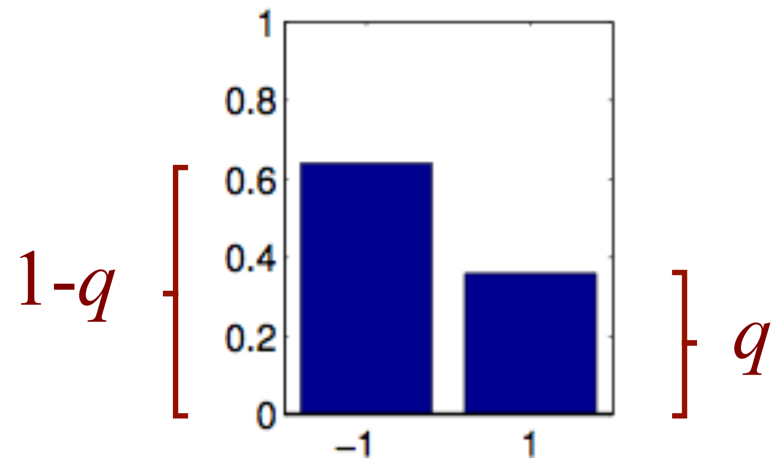
# Generative modeling

- The label selection is simply a biased coin flip (Bernoulli distribution)

indicator function: $\delta(a,b)=1$ if $a=b$
$\delta(a,b)=0$ if $a\neq b$

$$P(y;\theta) = q^{\delta(y,1)}(1-q)^{\delta(y,-1)}$$

where $q$ is included in $\theta$ (parameters that define the full distribution)



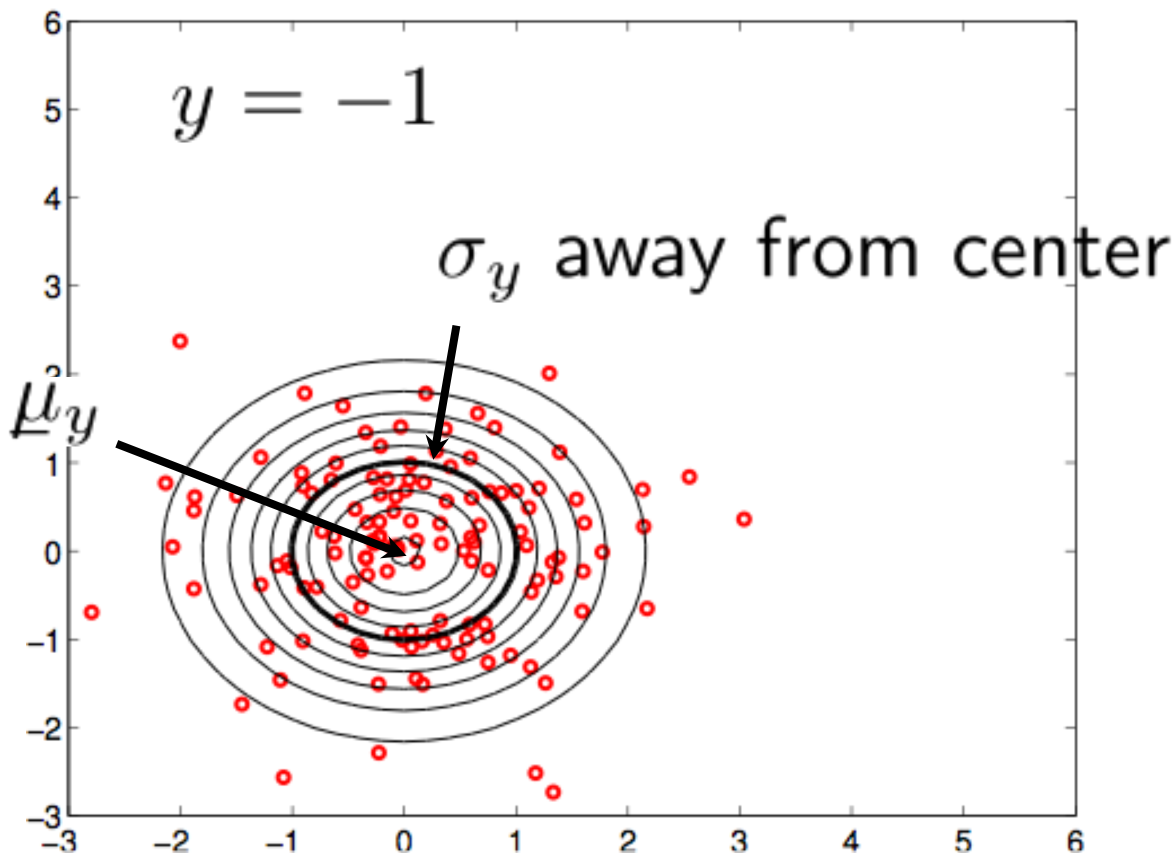
# Generative modeling

- We can use simple spherical Gaussian models for the class-conditional distributions

$\underline{x}$ and $\underline{\mu}_y$ are d-dimensional vectors

$$P(\underline{x}|y;\theta) = N(\underline{x};\, \underline{\mu}_y, \sigma_y^2 I)$$

$$= \frac{1}{(2\pi\sigma_y^2)^{d/2}} \exp\left\{ -\frac{1}{2\sigma_y^2} \|\underline{x} - \underline{\mu}_y\|^2 \right\}$$



# Generative modeling

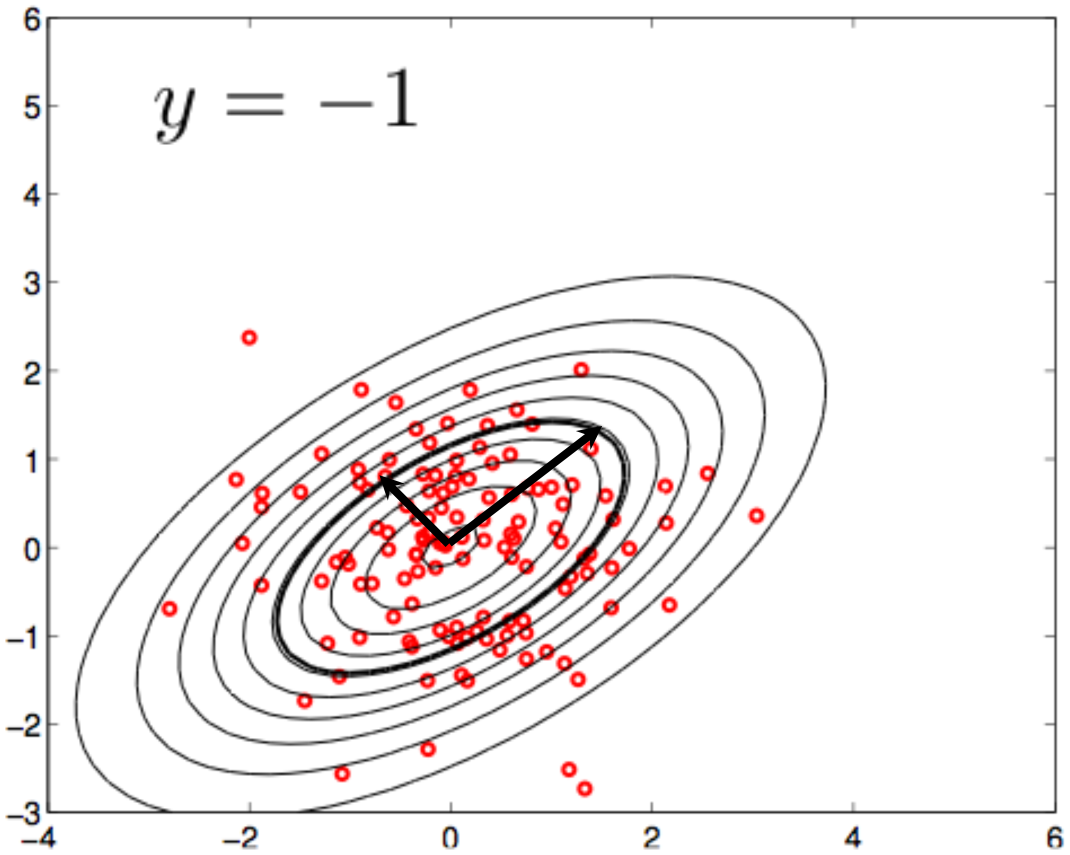
- We can also use full Gaussian models

$$P(\underline{x}|y;\theta) = N(\underline{x};\, \underline{\mu}_y, \Sigma_y)$$
$$= \frac{1}{(2\pi)^{d/2}|\Sigma_y|^{1/2}} \exp\left\{ -\frac{1}{2}(\underline{x} - \underline{\mu}_y)^T \Sigma_y^{-1} (\underline{x} - \underline{\mu}_y) \right\}$$

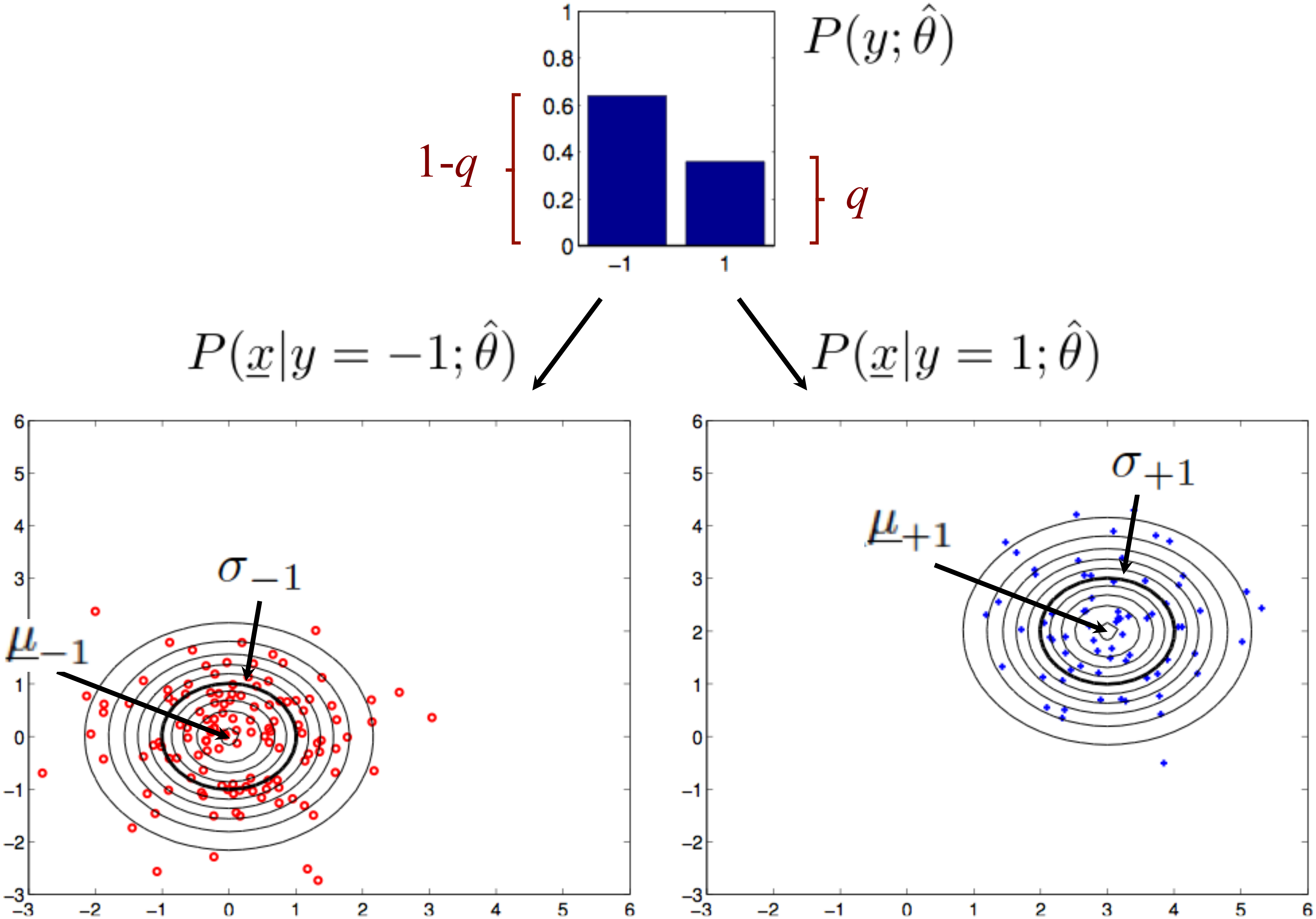
$$\Sigma_y = R \begin{bmatrix} \sigma_{y1}^2 & 0 \\ 0 & \sigma_{y2}^2 \end{bmatrix} R^T$$

rotation matrix

variances along the two principal axis



# Generative modeling: estimation



# Maximum likelihood estimation

- Our parameterized Gaussian model is

$$P(\underline{x}, y; \theta) = P(\underline{x}|y; \theta)P(y; \theta) = N(\underline{x};\, \mu_y, \sigma_y^2 I)\; q^{\delta(y,1)}(1-q)^{\delta(y,-1)}$$

- We find parameters $\theta = (\underline{\mu}_{+1}, \underline{\mu}_{-1}, \sigma^2_{+1}, \sigma^2_{-1}, q)$ that maximize the log-likelihood of the training data (examples and labels)

$$l(D; \theta) = \sum_{i=1}^{n} \log P(\underline{x}_i, y_i; \theta) = \sum_{i=1}^{n} \Big[ \log P(\underline{x}_i|y_i; \theta) + \log P(y_i; \theta) \Big]$$

See Lecture 11,
slide 20

# Maximum likelihood estimation

- Our parameterized Gaussian model is

$$P(\underline{x}, y; \theta) = P(\underline{x}|y;\theta)P(y;\theta) = N(\underline{x};\ \mu_y, \sigma_y^2 I)\ q^{\delta(y,1)}(1-q)^{\delta(y,-1)}$$

- We find parameters $\theta = (\underline{\mu}_{+1}, \underline{\mu}_{-1}, \sigma^2_{+1}, \sigma^2_{-1}, q)$ that maximize the log-likelihood of the training data (examples and labels)

$$l(D;\theta) = \sum_{i=1}^{n} \log P(\underline{x}_i, y_i; \theta) = \sum_{i=1}^{n} \Big[ \log P(\underline{x}_i|y_i;\theta) + \log P(y_i;\theta) \Big]$$

$$= \sum_{i=1}^{n} \Big[ -\frac{d}{2}\log(2\pi\sigma^2_{y_i}) - \frac{1}{2\sigma^2_{y_i}} \|\underline{x}_i - \mu_{y_i}\|^2 \Big] +$$

# Maximum likelihood estimation

- Our parameterized Gaussian model is

$$P(\underline{x}, y; \theta) = P(\underline{x}|y; \theta)P(y; \theta) = N(\underline{x};\ \mu_y, \sigma_y^2 I)\ q^{\delta(y,1)}(1-q)^{\delta(y,-1)}$$

- We find parameters $\theta = (\underline{\mu}_{+1}, \underline{\mu}_{-1}, \sigma_{+1}^2, \sigma_{-1}^2, q)$ that maximize the log-likelihood of the training data (examples and labels)

$$l(D; \theta) = \sum_{i=1}^{n} \log P(\underline{x}_i, y_i; \theta) = \sum_{i=1}^{n} \Big[ \log P(\underline{x}_i | y_i; \theta) + \log P(y_i; \theta) \Big]$$

$$= \sum_{i=1}^{n} \Big[ -\frac{d}{2} \log(2\pi\sigma_{y_i}^2) - \frac{1}{2\sigma_{y_i}^2} \|\underline{x}_i - \mu_{y_i}\|^2 \Big] +$$

$$+ \sum_{i=1}^{n} \Big[ \delta(y_i, 1) \log q + \delta(y_i, -1) \log(1-q) \Big]$$

indicator function

# Maximum likelihood estimation

$$l(D;\theta) = \sum_{i=1}^{n} \left[ -\frac{d}{2}\log(2\pi\sigma_{y_i}^2) - \frac{1}{2\sigma_{y_i}^2}\|\underline{x}_i - \mu_{y_i}\|^2 \right]$$
$$+ \sum_{i=1}^{n} \Big[ \delta(y_i, 1)\log q + \delta(y_i, -1)\log(1-q) \Big]$$

$$\frac{\partial}{\partial q} l(D;\theta) = \frac{\sum_{i=1}^{n} \delta(y_i, 1)}{q} - \frac{\sum_{i=1}^{n} \delta(y_i, -1)}{1-q} = 0$$

See Lecture 11, slide 23

# Maximum likelihood estimation

$$l(D;\theta) = \sum_{i=1}^{n} \left[ -\frac{d}{2}\log(2\pi\sigma_{y_i}^2) - \frac{1}{2\sigma_{y_i}^2}\|\underline{x}_i - \mu_{y_i}\|^2 \right] + \sum_{i=1}^{n} \Big[ \delta(y_i,1)\log q + \delta(y_i,-1)\log(1-q) \Big]$$

$$\frac{\partial}{\partial q} l(D;\theta) = \frac{\sum_{i=1}^{n}\delta(y_i,1)}{q} - \frac{\sum_{i=1}^{n}\delta(y_i,-1)}{1-q} = 0$$

$$\Rightarrow \hat{q} = \frac{1}{n}\sum_{i=1}^{n}\delta(y_i,1)$$

fraction of points labeled +1

# Maximum likelihood estimation

$$l(D;\theta) = \sum_{i=1}^{n} \left[ -\frac{d}{2}\log(2\pi\sigma_{y_i}^2) - \frac{1}{2\sigma_{y_i}^2}\|\underline{x}_i - \mu_{y_i}\|^2 \right]$$
$$+ \sum_{i=1}^{n} \left[ \delta(y_i, 1)\log q + \delta(y_i, -1)\log(1-q) \right]$$

$$\frac{\partial}{\partial \underline{\mu}_y} l(D;\theta) =$$

See Lecture 11, slide 23

# Maximum likelihood estimation

$$l(D;\theta) = \sum_{i=1}^{n} \left[ -\frac{d}{2}\log(2\pi\sigma_{y_i}^2) - \frac{1}{2\sigma_{y_i}^2}\|\underline{x}_i - \mu_{y_i}\|^2 \right] + \sum_{i=1}^{n} \left[ \delta(y_i, 1)\log q + \delta(y_i, -1)\log(1-q) \right]$$

$$\frac{\partial}{\partial \underline{\mu}_y} l(D;\theta) = \sum_{i=1}^{n} \delta(y, y_i) \frac{1}{\sigma_y^2}(\underline{x}_i - \mu_y) = 0$$

# Maximum likelihood estimation

$$l(D;\theta) = \sum_{i=1}^{n} \left[ -\frac{d}{2}\log(2\pi\sigma_{y_i}^2) - \frac{1}{2\sigma_{y_i}^2}\|\underline{x}_i - \mu_{y_i}\|^2 \right]$$

$$+ \sum_{i=1}^{n} \Big[ \delta(y_i, 1)\log q + \delta(y_i, -1)\log(1-q) \Big]$$

$$\frac{\partial}{\partial \underline{\mu}_y} l(D;\theta) = \sum_{i=1}^{n} \delta(y, y_i) \frac{1}{\sigma_y^2}(\underline{x}_i - \mu_y) = 0$$

$$\Rightarrow \hat{\underline{\mu}}_y = \frac{1}{\sum_{i=1}^{n}\delta(y, y_i)} \sum_{i=1}^{n} \delta(y, y_i)\, \underline{x}_i$$

average of points
in class y

# Maximum likelihood estimation

$$l(D;\theta) = \sum_{i=1}^{n} \left[ -\frac{d}{2}\log(2\pi\sigma_{y_i}^2) - \frac{1}{2\sigma_{y_i}^2}\|\underline{x}_i - \mu_{y_i}\|^2 \right]$$

$$+ \sum_{i=1}^{n} \left[ \delta(y_i,1)\log q + \delta(y_i,-1)\log(1-q) \right]$$

$$\frac{\partial}{\partial \sigma_y^2} l(D;\theta) = \sum_{i=1}^{n} \delta(y,y_i) \left[ -\frac{d}{2\sigma_y^2} + \frac{1}{2\sigma_y^4}\|\underline{x}_i - \hat{\mu}_y\|^2 \right] = 0$$

See Lecture 11, slide 23

# Maximum likelihood estimation

$$l(D;\theta) = \sum_{i=1}^{n} \left[ -\frac{d}{2}\log(2\pi\sigma_{y_i}^2) - \frac{1}{2\sigma_{y_i}^2}\|\underline{x}_i - \mu_{y_i}\|^2 \right]$$
$$+ \sum_{i=1}^{n} \left[ \delta(y_i,1)\log q + \delta(y_i,-1)\log(1-q) \right]$$

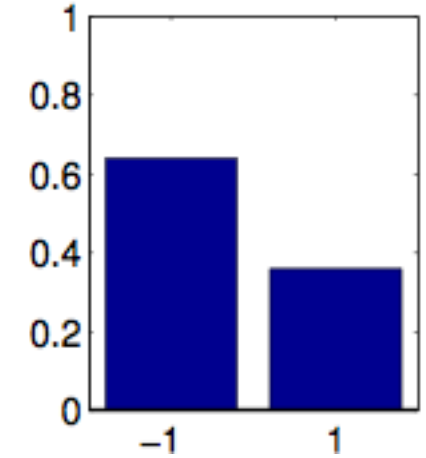
$$\frac{\partial}{\partial\sigma_y^2} l(D;\theta) = \sum_{i=1}^{n} \delta(y,y_i)\left[ -\frac{d}{2\sigma_y^2} + \frac{1}{2\sigma_y^4}\|\underline{x}_i - \hat{\mu}_y\|^2 \right] = 0$$

$$\Rightarrow \hat{\sigma}_y^2 = \frac{1}{d\sum_{i=1}^{n}\delta(y,y_i)} \sum_{i=1}^{n} \delta(y,y_i)\,\|\underline{x}_i - \hat{\mu}_y\|^2$$

average per dimension squared
error in class y

# Generative modeling: classification

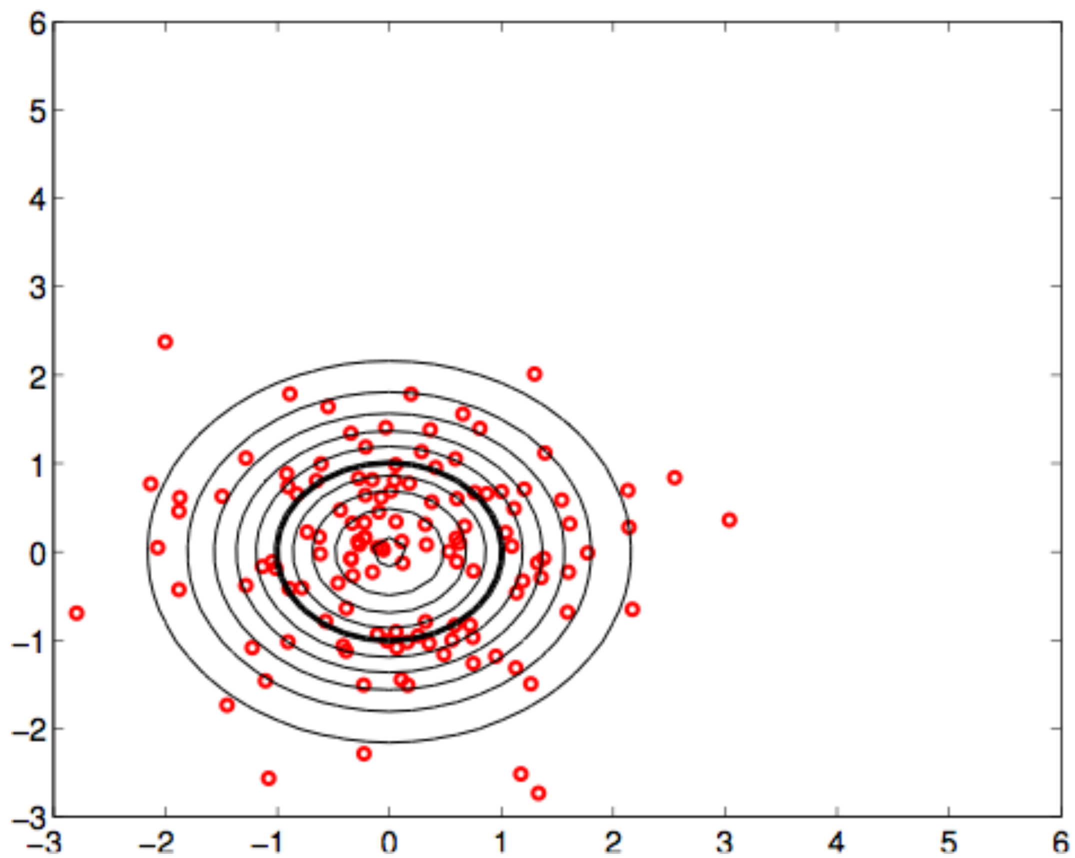
$$\theta = (\underline{\mu}_{+1}, \underline{\mu}_{-1}, \sigma^2_{+1}, \sigma^2_{-1}, q)$$

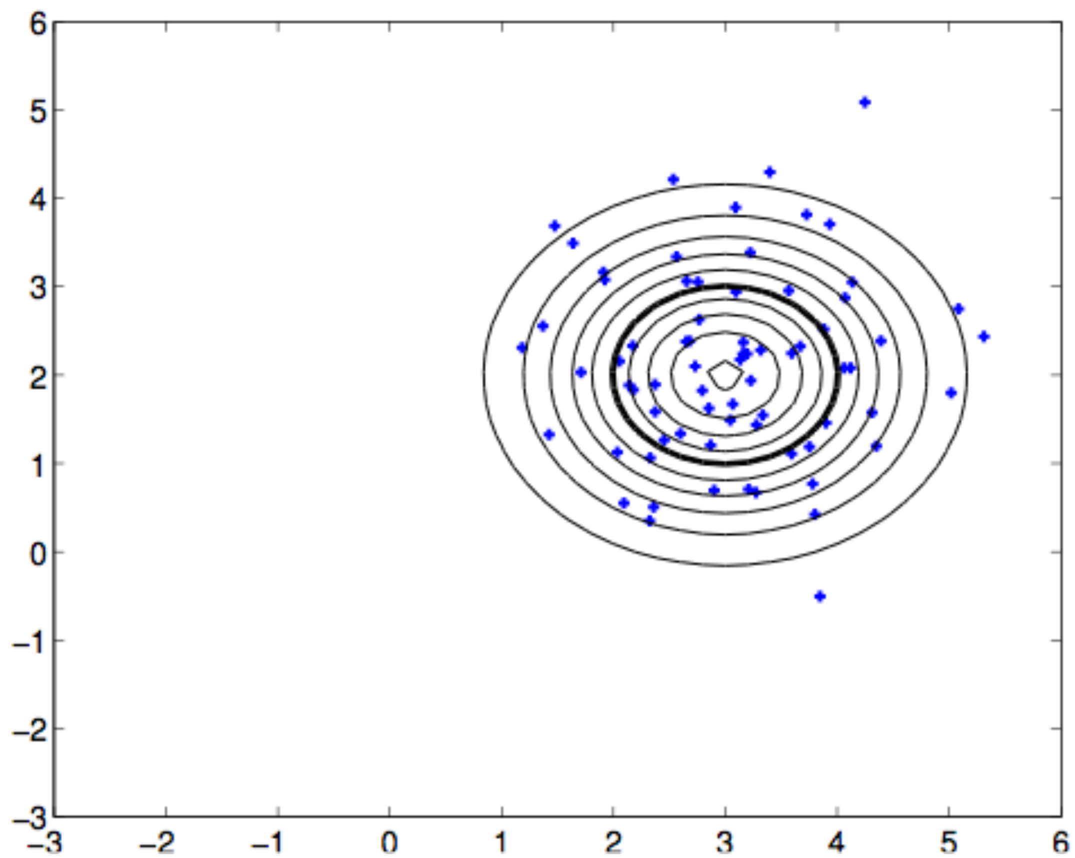



$P(y; \hat{\theta})$

$P(\underline{x}|y = -1; \hat{\theta})$

$P(\underline{x}|y = 1; \hat{\theta})$



# Decision boundary

- Given $\underline{x}$, predict the label ($+1$ or -$1$) with highest probability
  - Predict label $y = +1$ if $P(y = 1 | \underline{x}; \hat{\theta}) > P(y = -1 | \underline{x}; \hat{\theta})$

    **by conditional probability** $\dfrac{P(\underline{x}, y = 1; \hat{\theta})}{P(\underline{x}; \hat{\theta})} > \dfrac{P(\underline{x}, y = -1; \hat{\theta})}{P(\underline{x}; \hat{\theta})}$

    **equivalent to** $P(\underline{x}, y = 1; \hat{\theta}) > P(\underline{x}, y = -1; \hat{\theta})$

  - Predict label $y =$ -$1$ if $P(\underline{x}, y = 1; \hat{\theta}) < P(\underline{x}, y = -1; \hat{\theta})$

- The decision boundary is the set of $\underline{x}$ for which we do not know what label ($+1$ or -$1$) to predict

$$P(\underline{x}, y = 1; \hat{\theta}) = P(\underline{x}, y = -1; \hat{\theta})$$

**or** $\dfrac{P(\underline{x}, y = 1; \hat{\theta})}{P(\underline{x}, y = -1; \hat{\theta})} = 1$ **or** $\log \dfrac{P(\underline{x}, y = 1; \hat{\theta})}{P(\underline{x}, y = -1; \hat{\theta})} = 0$

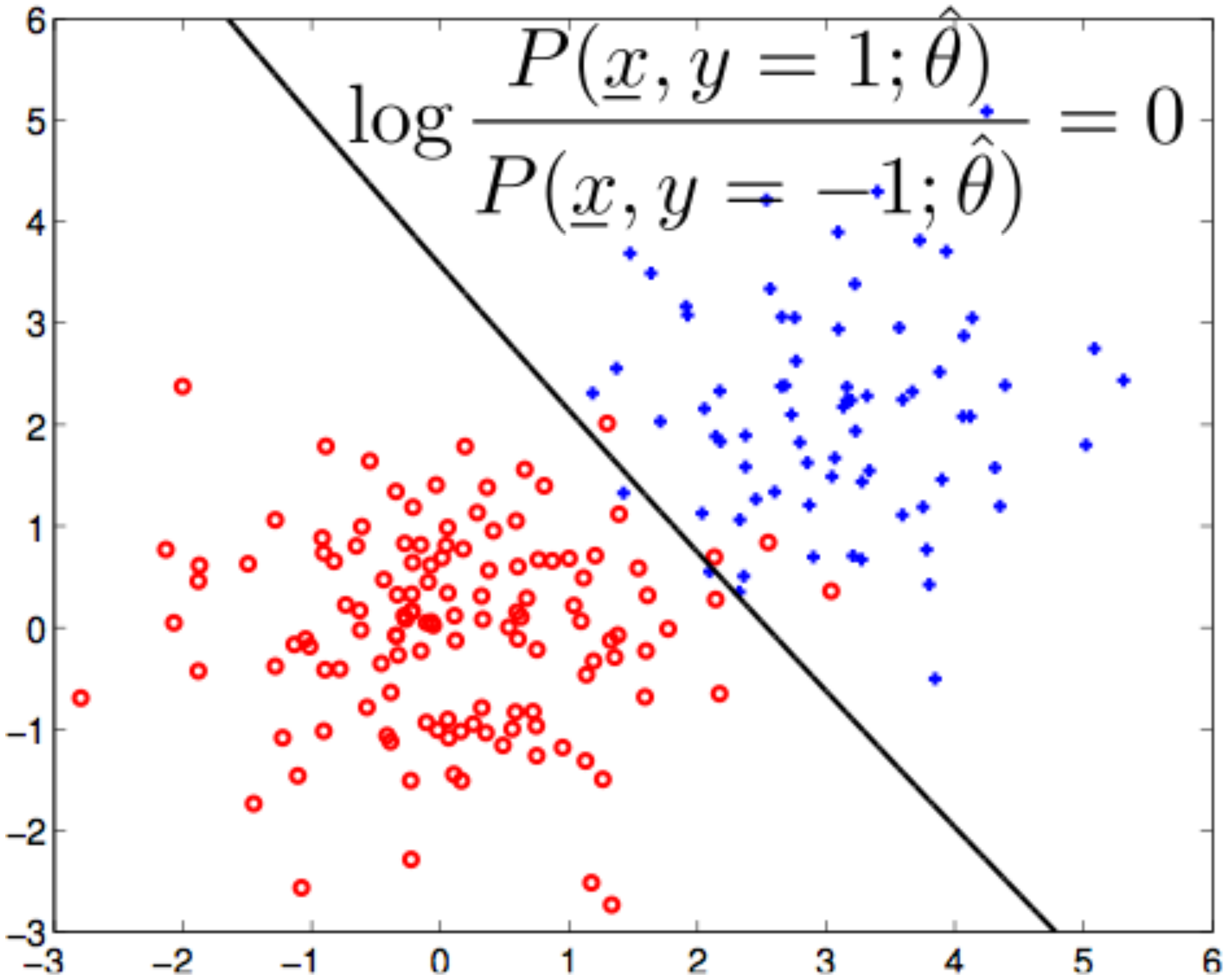
# Decision boundary

- The resulting decision boundary corresponds to all $\underline{x}$ such that

$$\begin{aligned}\log \frac{P(\underline{x}, y=1; \hat{\theta})}{P(\underline{x}, y=-1; \hat{\theta})} &= \log \frac{P(y=1; \hat{\theta})}{P(y=-1; \hat{\theta})} + \log \frac{P(\underline{x}|y=1; \hat{\theta})}{P(\underline{x}|y=-1; \hat{\theta})} \\ &= \log \frac{\hat{q}}{1-\hat{q}} - \frac{d}{2} \log(\frac{\hat{\sigma}^2_{+1}}{\hat{\sigma}^2_{-1}}) \\ &\quad - \frac{1}{2\hat{\sigma}^2_{+1}} \|\underline{x} - \underline{\hat{\mu}}_{+1}\|^2 + \frac{1}{2\hat{\sigma}^2_{-1}} \|\underline{x} - \underline{\hat{\mu}}_{-1}\|^2 \\ &= 0\end{aligned}$$

- This is linear in $\underline{x}$ if $\hat{\sigma}^2_{+1} = \hat{\sigma}^2_{-1}$ (otherwise quadratic)

# Decision boundary




- This is very close to the optimal since the data were generated from two Gaussians with the same variance

# Probability predictions

- The model also permits us to evaluate probabilities over the possible class labels such as

$$P(y = 1|\underline{x}; \hat{\theta}) = \frac{P(\underline{x}, y = 1; \hat{\theta})}{\sum_{y' \in \{-1,1\}} P(\underline{x}, y'; \hat{\theta})}$$

the denominator is: $P(\underline{x};\theta)$