So far, supervised learning : $h: x \to \mathbb{R}$ ← regression
$h: x \to \{0, 1, \ldots, k\}$ ← classification

Unsupervised learning

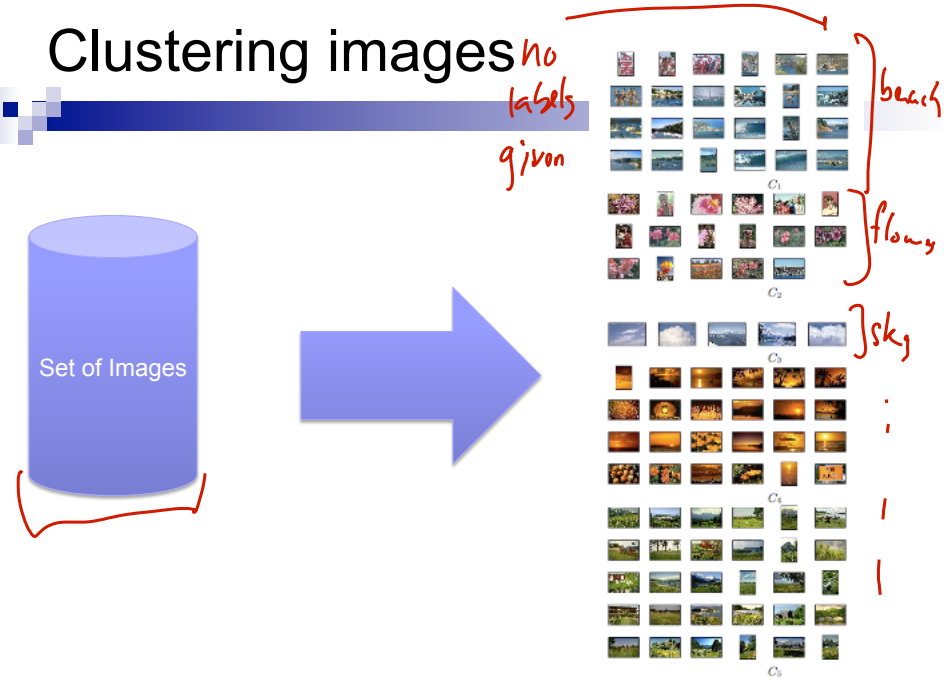
Clustering
K-means Continued

Machine Learning – CSE446
Carlos Guestrin
University of Washington

May 15, 2013

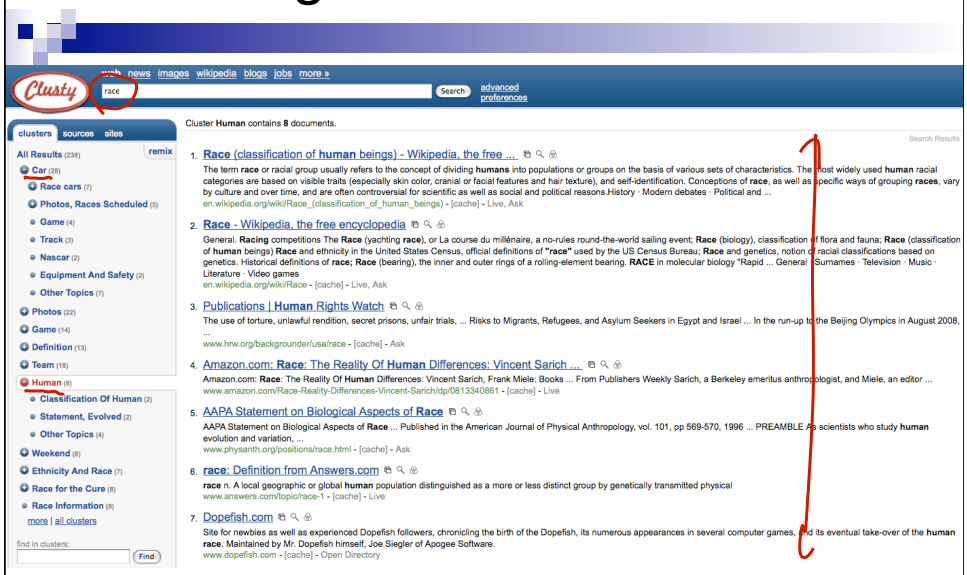
Clustering images

no labels given

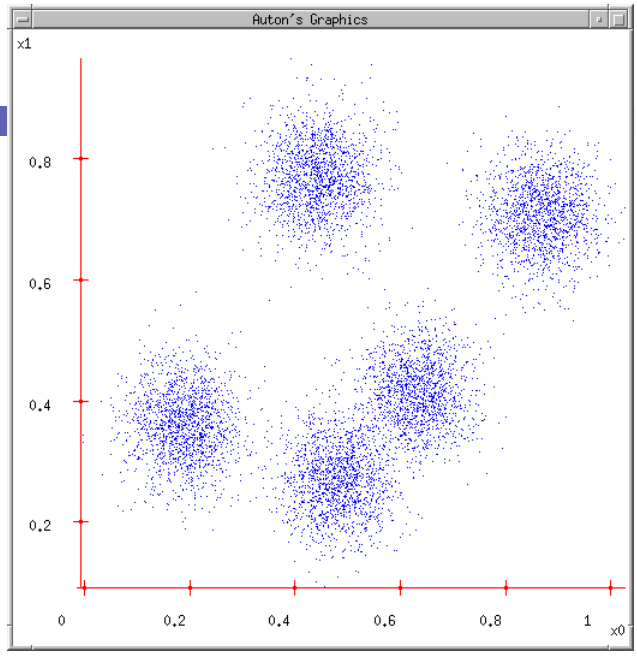


[Goldberger et al.]

Clustering web search results

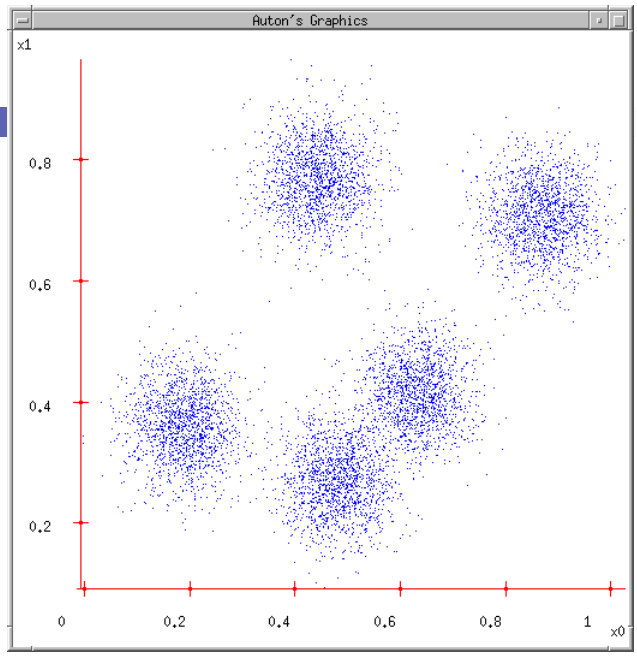


Some Data



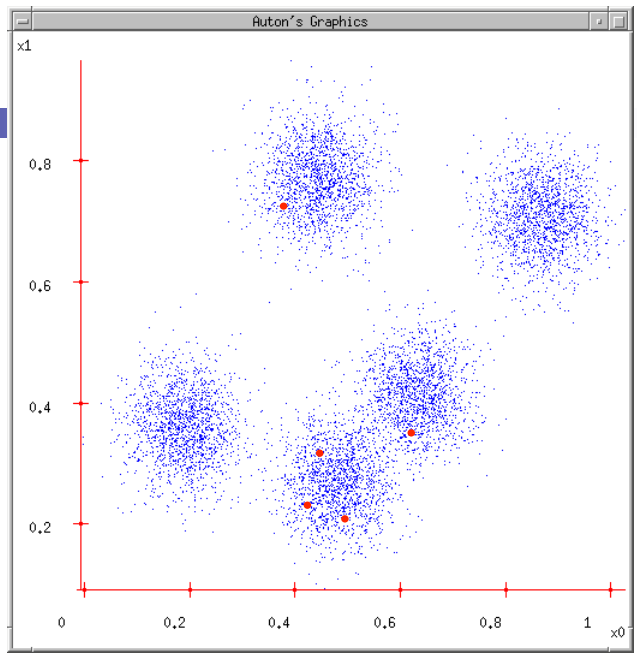
K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*



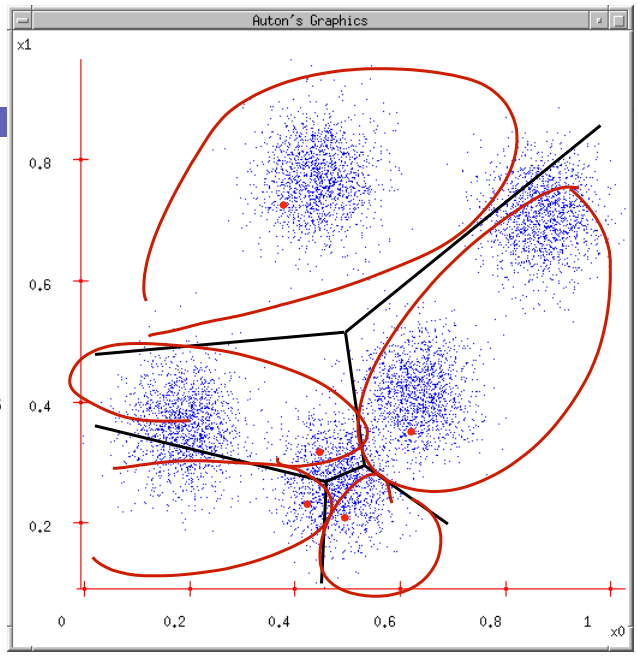
K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations



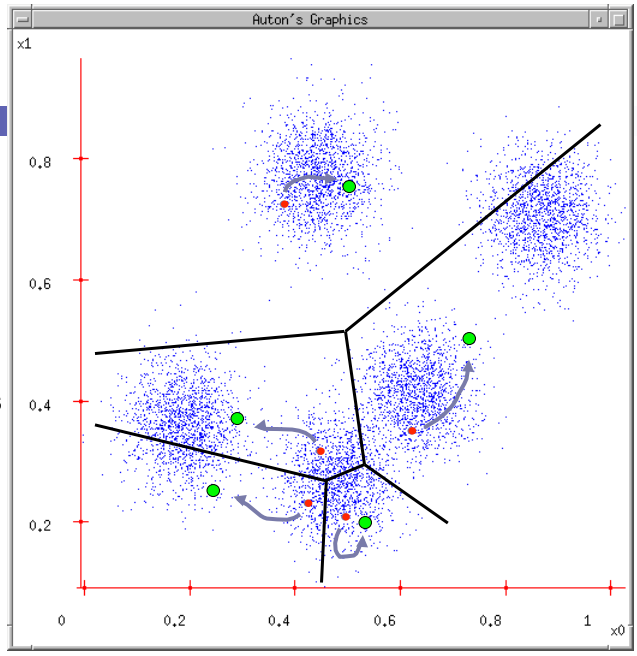
K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to. (Thus each Center "owns" a set of datapoints)



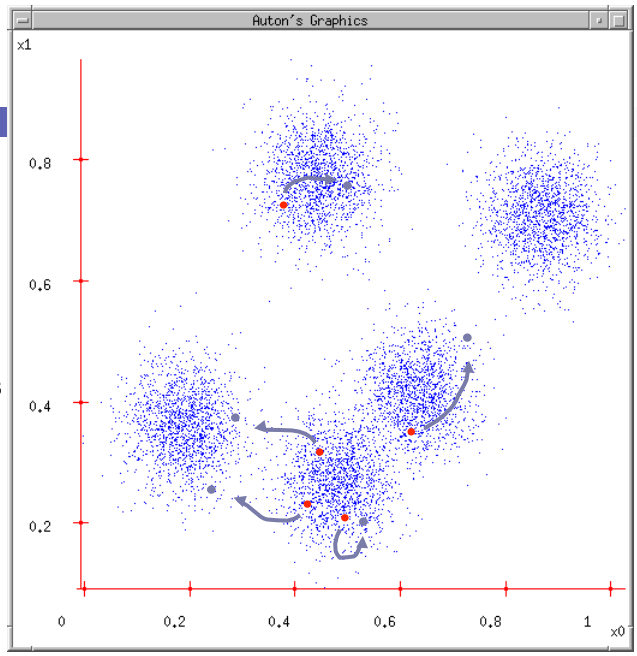
K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns



K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns…
5. …and jumps there
6. …Repeat until terminated!



K-means

- Randomly initialize *k* centers or smartly
  - $\mu^{(0)} = \mu_1^{(0)},\ldots, \mu_k^{(0)}$

  converge when nothing moves (no point changes its cluster)
- **Classify**: Assign each point $j \in \{1,\ldots n\}$ to nearest center:
  - $C^{(t)}(j) \leftarrow \arg\min_i ||\mu_i - x_j||^2$

  ith cluster center; data point; fix μ, OPTC
- **Recenter**: μ_i becomes centroid of its point:
  - $\mu_i^{(t+1)} \leftarrow \arg\min_\mu \sum_{j:C(j)=i} ||\mu - x_j||^2$ ← $\mu_i = \frac{\sum_{j:C(j)=i} x_j}{\text{num of points assigned to cluster } i}$
  - Equivalent to μ_i ← average of its points!

What is K-means optimizing?

- Potential function F(μ,C) of centers μ and point allocations C:

  distance over all points

  - $F(\mu, C) = \sum_{j=1}^{N} ||\mu_{C(j)} - x_j||^2$

  k centers allocations

- Optimal K-means:
  - $\min_\mu \min_C F(\mu,C)$

  coordinate descent

  optimization is hard

$\min_a \min_b f(a,b) = \min_a \min_b g(a) + h(b) = [\min_a g(a)] + [\min_b h(b)]$ $f(a,b) = g(a) + h(b)$

Does K-means converge??? Part 1

- Optimize potential function:

$$\min_\mu \min_C F(\mu, C) = \min_\mu \min_C \sum_{i=1}^{k} \sum_{j:C(j)=i} ||\mu_i - x_j||^2$$

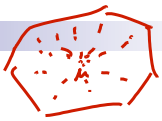
- Fix μ, optimize C

$$\min_{C: C(1)\ldots C(N)} \sum_{j=1}^{N} ||\mu_{C(j)} - x_j||^2 = \min_{C(1)} \min_{C(2)} \cdots \min_{C(N)} \sum_{j=1}^{N} ||\mu_{C(j)} - x_j||^2$$

$$= \sum_{j=1}^{N} \underbrace{\min_{C(j)} ||\mu_{C(j)} - x_j||^2}_{\text{exactly the "classification step"}}$$

indep. minimizations

Does K-means converge??? Part 2



- Optimize potential function:

$$\min_{\mu}\min_{C} F(\mu, C) = \min_{\mu}\min_{C} \sum_{i=1}^{k} \sum_{j:C(j)=i} ||\mu_i - x_j||^2$$

- Fix C, optimize μ ← recenter

$$\min_{\mu:\mu_1,\dots,\mu_k} \sum_{j=1}^{N} \|\mu_{C(j)} - x_j\|^2 = \min_{\mu:\mu_1,\dots,\mu_k} \sum_{i=1}^{k} \sum_{j:C(j)=i} \|\mu_i - x_j\|^2$$

$$= \sum_{i=1}^{k} \underbrace{\min_{\mu_i} \sum_{j:C(j)=i} \|\mu_i - x_j\|^2}$$

μ_i = center of points = average = $\dfrac{\sum_{j:C(j)=i} x_j}{\text{number of points in cluster } i}$

Coordinate descent algorithms

$$\min_{\mu}\min_{C} F(\mu, C) = \min_{\mu}\min_{C} \sum_{i=1}^{k} \sum_{j:C(j)=i} ||\mu_i - x_j||^2 \geq 0$$

- Want: $\min_a \min_b F(a,b)$
- Coordinate descent:
  - fix a, minimize b
  - fix b, minimize a
  - repeat
- Converges!!!
  - if F is bounded below
  - to a (often good) local optimum
    - as we saw in applet (play with it!)
      - (For LASSO it converged to the optimum)

a,b · $F(a,b)$

$\downarrow \min_a$

a',b $F(a',b) \leq F(a,b)$

$\downarrow \min_b$

a',b' $F(a',b') \leq F(a',b) \leq F(a,b)$

$\vdots$

non-increasing sequence

lower bound

Random restarts help

- K-means is a coordinate descent algorithm!

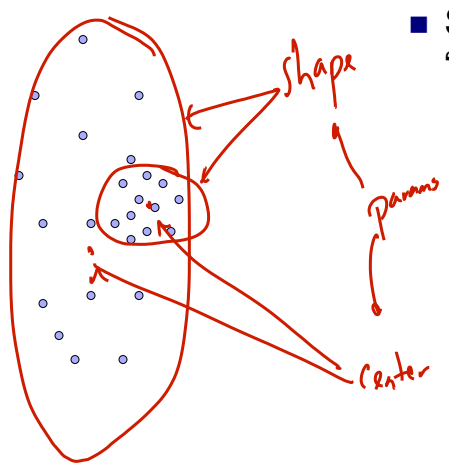
Mixtures of Gaussians

Machine Learning – CSE446
Carlos Guestrin
University of Washington

May 15, 2013

(One) bad case for k-means

- Clusters may overlap
- Some clusters may be "wider" than others



Gaussians in *m* Dimensions

Covariance matrix Σ is a bga

m-dim

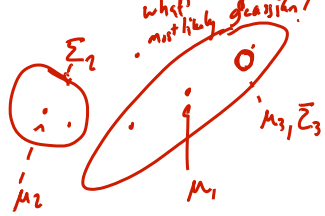
$$P(\mathbf{x}) = \frac{1}{(2\pi)^{m/2} \, \|\Sigma\|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)\right]$$

mean ← m-dim vector

1. D : $p(x) \propto e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

shape Σ_1

what's most likely gaussian?

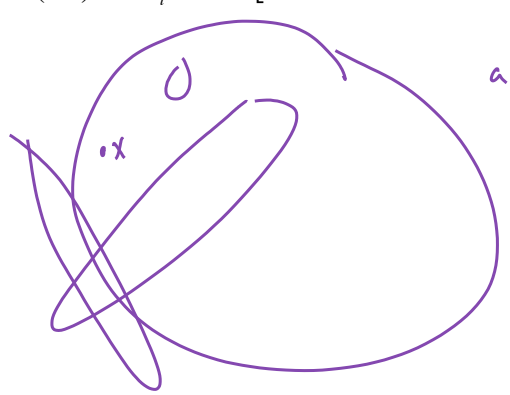


Suppose You Have a Gaussian For Each Class

$\|a-b\|_\Sigma = (a-b)\Sigma(a-b)$

$$P(\mathbf{x} \mid y = i) \propto \frac{1}{(2\pi)^{m/2} \, \|\Sigma_i\|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu_i)^T \Sigma_i^{-1} (\mathbf{x}-\mu_i)\right]$$

$P(y=i \mid x)$



a distance metric per class

Gaussian Bayes Classifier

- You have a Gaussian over **x** for each class y=i:

$$P(\mathbf{x} \mid y=i) = \frac{1}{(2\pi)^{m/2} |\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu_i)^T \Sigma_i^{-1} (\mathbf{x}-\mu_i)\right]$$

- But you need probability of class y=i given **x**:

$P(Y=i \mid X)$

- Thank you Bayes Rule!!

cluster data likelihoods

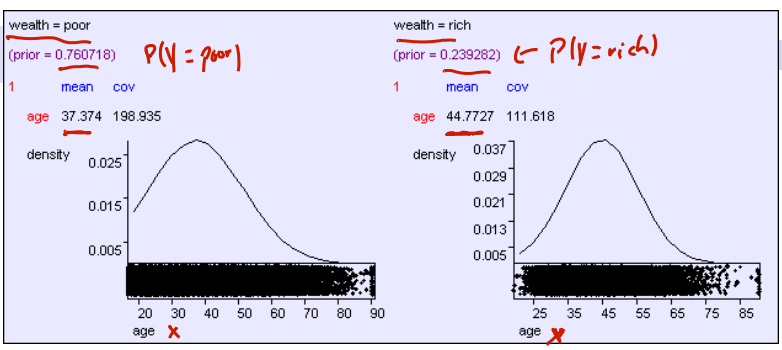
prior: before we see x, what fraction of points we expect to fall in cluster i

$$P(y=i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid y=i) P(y=i)}{p(\mathbf{x})}$$

normalizer

$$\propto \frac{1}{(2\pi)^{m/2} |\Sigma_i|^{1/2}} \exp -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i) \cdot P(Y=i)$$

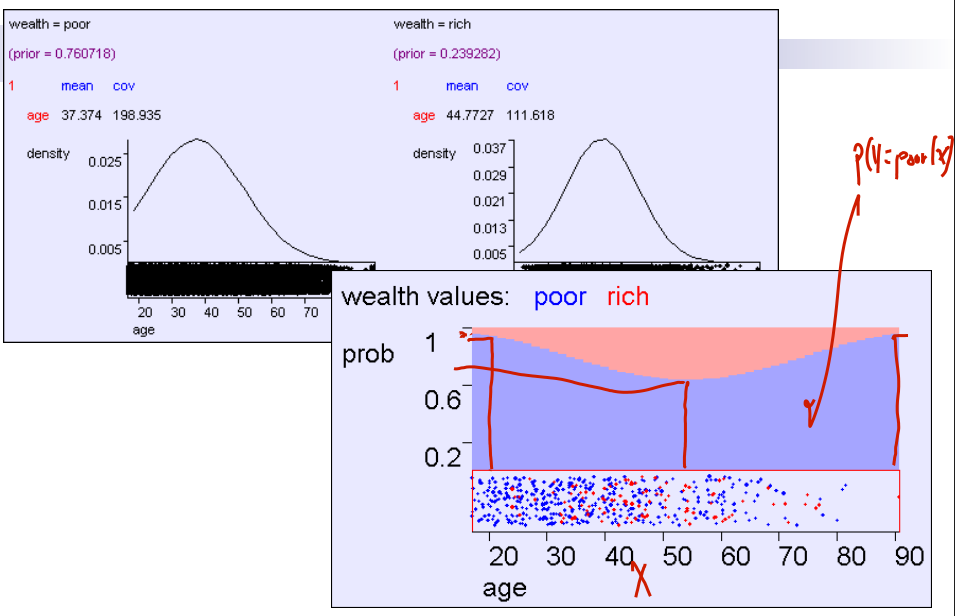
Predicting wealth from age



$P(Y=\text{poor} \mid X)$

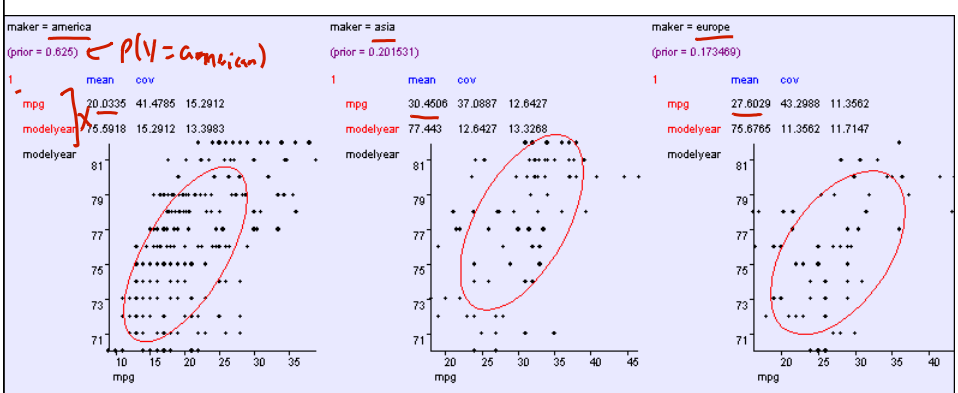
$= P(Y=\text{poor} \mid \text{age})$

Predicting wealth from age



Learning modelyear , mpg ---> maker

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots & \sigma_m^2 \end{pmatrix}$$



General: $O(m^2)$ parameters

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots & \sigma_m^2 \end{pmatrix}$$

classification $\quad \arg\max_y P(Y=y \mid X)$