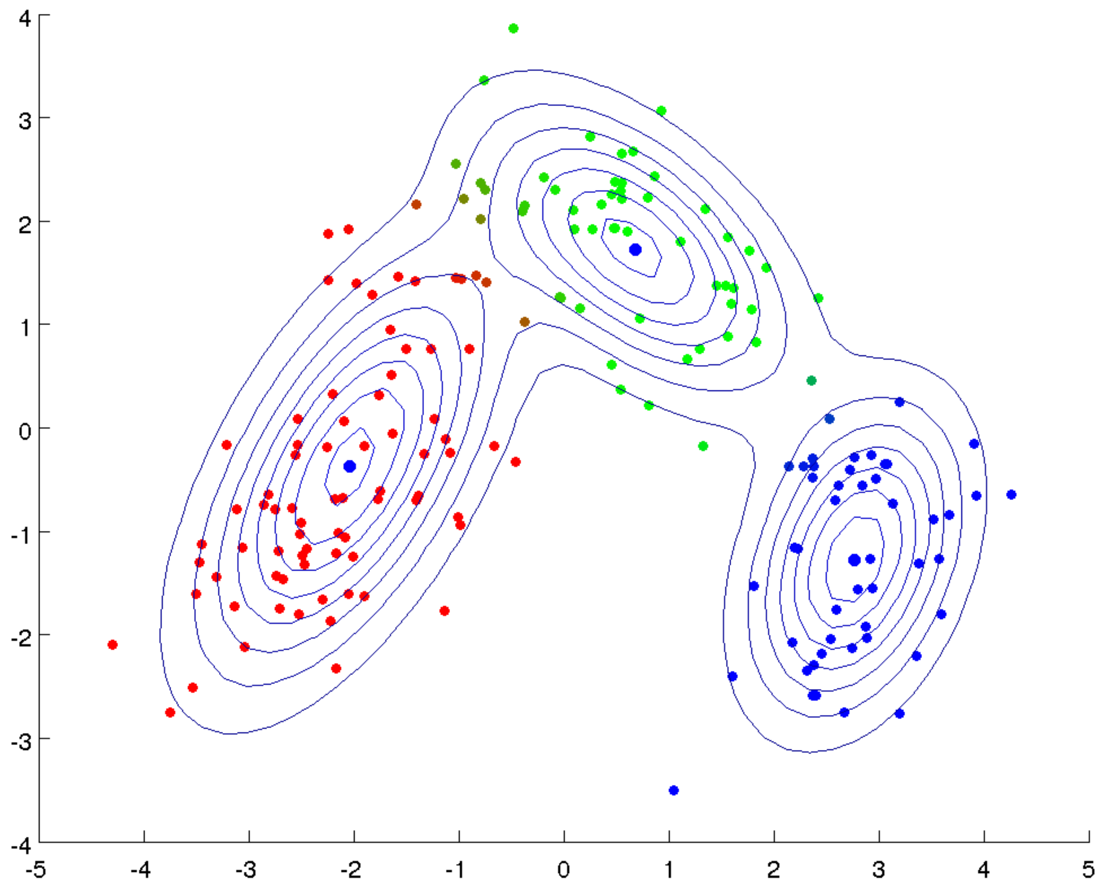
# Feature Selection & Midterm review

Le Song

Machine Learning
CSE/ISYE 6740, Fall 2019

# Wine dataset

- First run PCA to reduce the dimension to 2
- Clear cluster structure
- Can we fit 3 Gaussians?



# Theory underlying EM

- Recall that in MLE, we intend to learn the model parameter that would have maximize the likelihood of the data.
  - $l(\theta; D) = \log \sum_z p(x, z|\theta) = \log \sum_z p(x|\, z, \theta) P(z|\theta)$

- But we are iterating these:
  - Expectation step (E-step)
    - $f(\theta) = E_{q(z)}[\log p(x, z|\theta)], where\ q(z) = P(z|x, \theta^t)$
  - Maximization step (M-step)
    - $\theta^{t+1} = argmax_\theta\ \ f(\theta)$

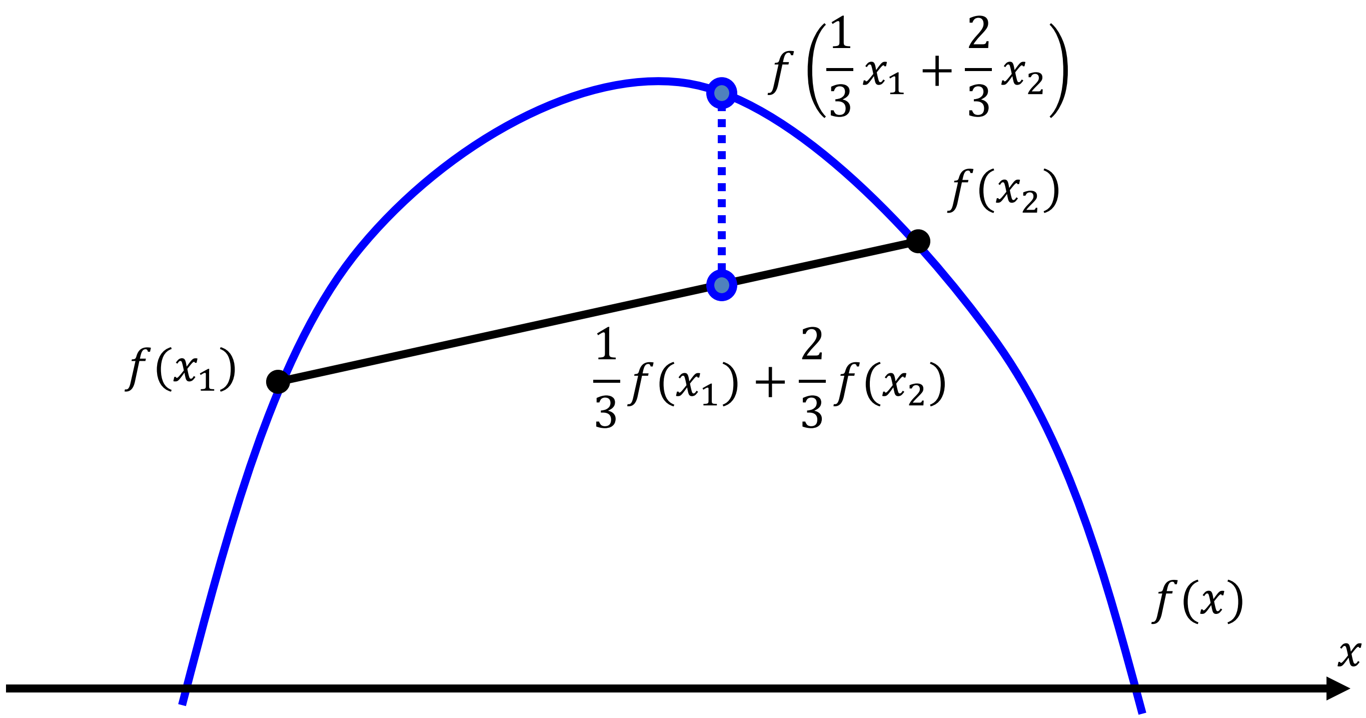
- Does maximizing this surrogate yield a maximizer of the likelihood?

# Jensen's inequality

- For concave function $f(x)$, eg. log(x)
  - $f(\sum_i \alpha_i x_i) \geq \sum_i \alpha_i f(x_i)\,, where\ \sum_i \alpha_i = 1, \alpha_i \geq 0$

- Most general case: If $x$ is a random variable, and $f$ is concave,

$$f(\mathbf{E}x) \geq \mathbf{E}f(x)$$



# Lower bound of log-likelihood

- Log-likelihood $l(x;\theta) = \log \sum_z p(x,z|\theta)$

$$= \log \sum_z q(z) \frac{p(x,z|\theta)}{q(z)} \quad (\text{arbitrary } q(z))$$

$$\geq \sum_z q(z) \log \frac{p(x,z|\theta)}{q(z)} \quad (Jensen's\ inequality\ f\left(\sum_i \alpha_i x_i\right) \geq \sum_i \alpha_i f(x_i))$$

$$= \sum_z q(z) \log p(x,z|\theta) - \sum_z q(z) \log q(z)$$

$$= E_{q(z)}[\log p(x,z|\theta)] + H_{q(z)}$$

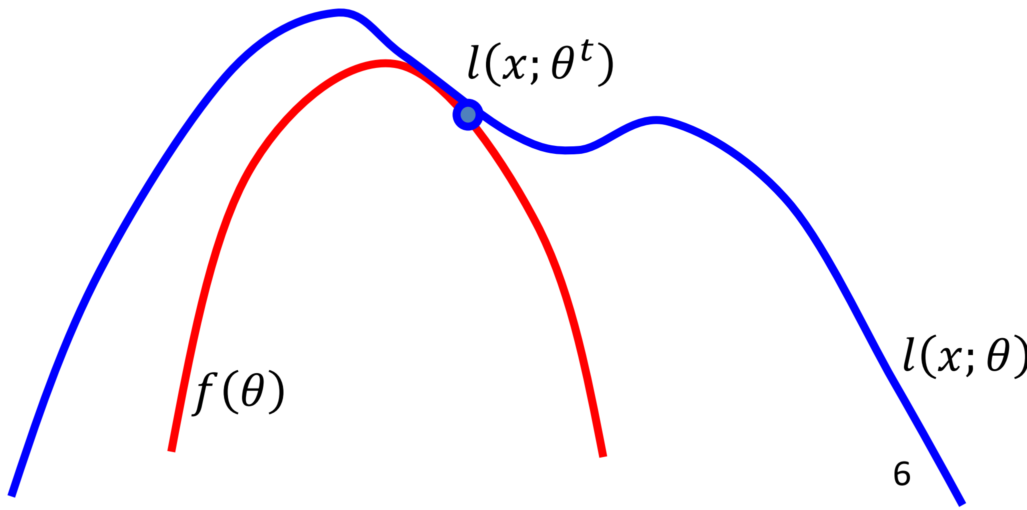
What $q$ to use?

# What attains equality?

- $q(z) = p(z|x, \theta^t)$: posterior of $z$ given $x$ attains the equality at $\theta^t$

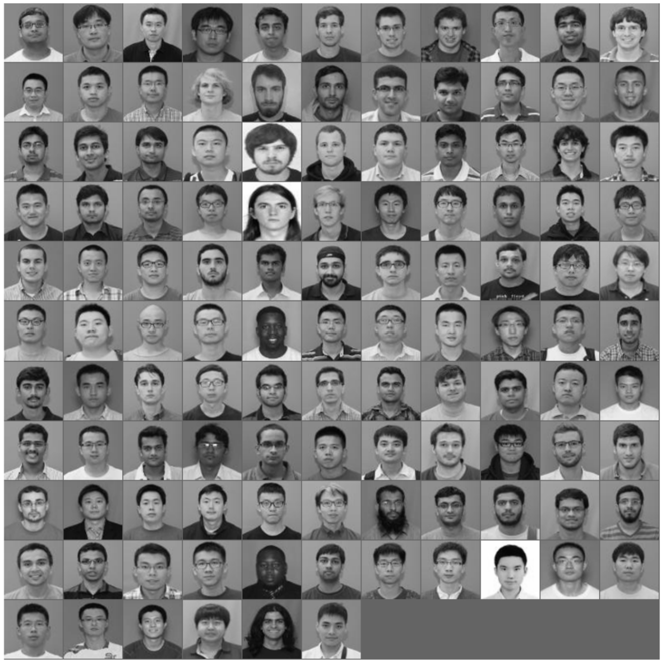
- Let $F(q, \theta) = \sum_z q(z) \log \frac{p(x, z|\theta)}{q(z)} \leq l(x; \theta) = \log \sum_z p(x, z|\theta)$

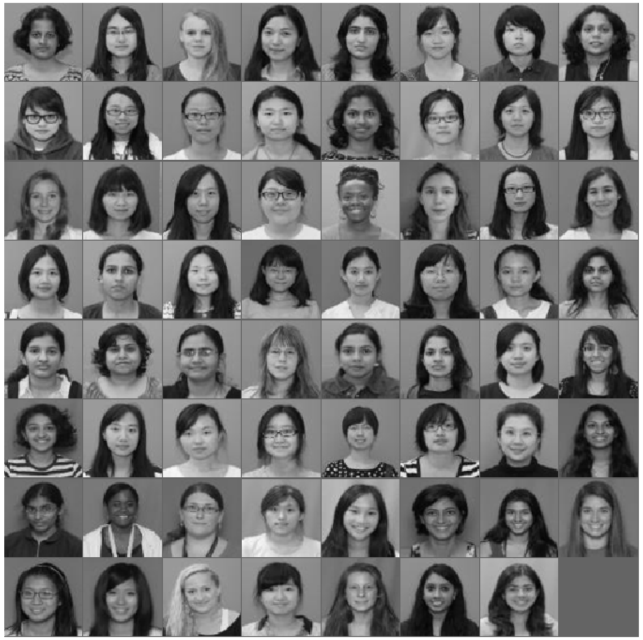
- $F(p(z|x, \theta^t), \theta^t) = \sum_z p(z|x, \theta^t) \log \frac{p\left(x, z \middle| \theta^t\right)}{p\left(z \middle| x, \theta^t\right)}$
- $= \sum_z p(z|x, \theta^t) \log p(x|\theta^t)$
- $= \log p(x|\theta^t)$
- $= \log \sum_z p(x, z|\theta^t)$



# Feature selection

- What are the best pixels for classifying photos of boys and girls?





# A feature selection algorithm

- Given a dataset $S = \{(x^1, y^1), \dots, (x^m, y^m)\},\ x \in R^d, y = \{1, \dots, K\}$ Label: male, female…
- For each value of the label $y = k$
  - Estimate density $p(y = k)$
- For each feature $x_i$ 下标：dimension; 上标：数据
  - Estimate its density $p(x_i)$
  - For each value of the label $y = k$
    - Estimate the density $p(x_i | y = k)$
  - Score feature $x_i$ using

$$I_i = \int \sum_{k=1}^{K} p(x_i | y = k) p(y = k) \log_2 \frac{p(x_i | y = k)}{p(x_i)} \ dx_i$$

- Choose those feature $x_i$ with high score $I_i$

# Informativeness of a feature

- We are uncertain about the label $Y$ before seeing any input
  - Suppose we quantify using $H(Y)$

- Given a particular feature $X_i$, the uncertainty of $Y$ changes
  - Suppose we quantify using $H(Y|X_i)$

- The reduction in uncertainty is the informativeness of feature $X_i$
  - $I(X_i, Y) = \ H(Y) - \ H(Y|X_i)$

- How to quantify uncertainty?

# Entropy: quantify uncertainty

- Entropy $H(Y)$ of a random variable $Y$

$$H(Y) = -\sum_{k=1}^{K} P(y=k)\log_2 P(y=k)$$

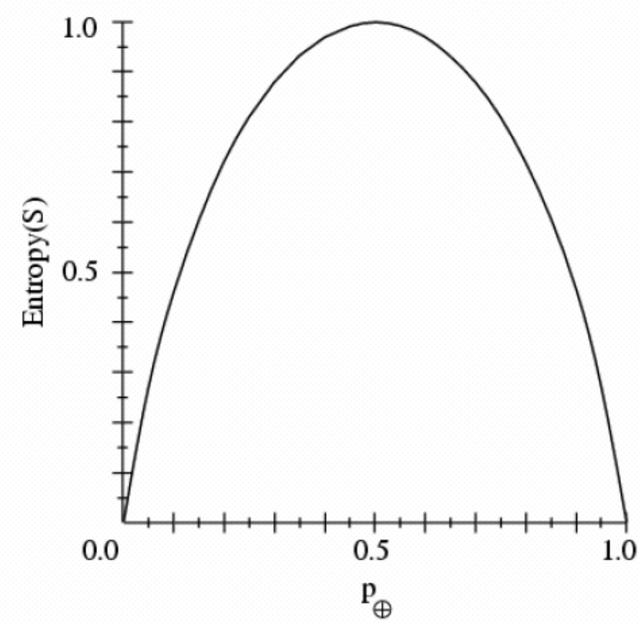
- $H(Y)$ is the expected number of bits needed to encode a randomly drawn value of $Y$ (under most efficient code)

- Information theory:

  Most efficient code assigns $-\log_2 P(Y=k)$ bits to encode the message $Y=k$, So, expected number of bits to code one random $Y$ is:

$$-\sum_{k=1}^{K} P(y=k)\log_2 P(y=k)$$

# Sample Entropy




- $S$ is a sample of coin flips
- $p_+$ is the proportion of heads in $S$
- $p_-$ is the proportion of tails in $S$
- Entropy measure the uncertainty of $S$

$$H(S) \equiv -p_+ \log_2 p_+ - p_- \log_2 p_-$$

# Examples for computing Entropy

$$H(S) \equiv -p_{+} \log_2 p_{+} - p_{-} \log_2 p_{-}$$

| | |
|---|---|
| head | **0** |
| tail | **6** |

**P(h) = 0/6 = 0     P(t) = 6/6 = 1**

**Entropy = – 0 log 0 – 1 log 1 = – 0 – 0 = 0**

| | |
|---|---|
| head | **1** |
| tail | **5** |

**P(h) = 1/6     P(t) = 5/6**

**Entropy = – (1/6) $\log_2$ (1/6) – (5/6) $\log_2$ (1/6) = 0.65**

| | |
|---|---|
| head | **2** |
| tail | **4** |

**P(h) = 2/6     P(t) = 4/6**

**Entropy = – (2/6) $\log_2$ (2/6) – (4/6) $\log_2$ (4/6) = 0.92**

# Conditional entropy

- Conditional entropy $H(Y|X)$ of a random variable $Y$ given $X_i$

**H(y|x) = - \sum_i \sum_j p(x_i, y_j) log p(y_j |x_i)**

$$H(Y|X_i) = -\int \left( \sum_{k=1}^{K} P(y=k|x_i) \log_2 P(y=k) \right) p(x_i) dx_i$$

**H(x,y) = - \sum_i \sum_j p(x,y) log p(x,y) = H(x)+H(y|x)**

- Quantify the uncertainty in $Y$ after seeing feature $X_i$

- $H(Y)$ is the expected number of bits needed to encode a randomly drawn value of $Y$
  - given $X_i$, and
  - average over the likelihood of seeing particular value of $x_i$

# Mutual information: reduction in uncertainty

- Mutual information: quantify the reduction in uncerntainty in $Y$ after seeing feature $X_i$

$$I(X_i, Y) = H(Y) - H(Y|X_i)$$

- The more the reduction in entropy, the more informative a feature.

- Mutual information is symmetric
  - $I(X_i, Y) = I(Y, X_i) = H(X_i) - H(X_i|Y)$
  - $I(Y, X_i) = \int \sum_k^K p(x_i, y = k) \log_2 \frac{p(x_i, y=k)}{p(x_i)p(y=k)} \; dx_i$
  - $= \int \sum_k^K p(x_i|y = k)p(y = k) \log_2 \frac{p(x_i|y = k)}{p(x_i)} \; dx_i$

# A feature selection algorithm

- Given a dataset $S = \{(x^1, y^1), \ldots, (x^m, y^m)\},\ x \in R^d, y = \{1, \ldots, K\}$
- For each value of the label $y = k$
  - Estimate density $p(y = k)$
- For each feature $x_i$
  - Estimate its density $p(x_i)$
  - For each value of the label $y = k$
    - Estimate the density $p(x_i | y = k)$
  - Score feature $x_i$ using $I_i = \int \sum_{k=1}^{K} p(x_i | y = k) p(y = k) \log_2 \frac{p\left(x_i \middle| y = k\right)}{p(x_i)}\ dx_i$
- Choose those feature $x_i$ with high score $I_i$

# Midterm Review

# Keys topics before midterm

- Unsupervised learning techniques
  - Dimensionality reduction
    - PCA
    - Graph based methods
  - Clustering
    - Kmeans
    - Graph based methods (spectral algorithms)
  - Density estimation
    - Parametric models
    - Histogram
    - Kernel density estimator
    - Mixture of Gaussian
  - Feature selection

# The process of designing ML systems

- What is the objective?
  - Extract group? Visualization? Reduce computation/memory? Compress data? Find useful features? Classification?
- Formulate the objective
  - Understand your data, and make assumptions: Independent? variance enough? Linear? Gaussian? Euclidean distance?
  - Parametrization: parametric? Nonparametric? Prior? Constraint?
- Looking for algorithms
  - Convex? Nonconvex? Computational and memory complexity? Iterative or one-shot? Global best? Guarantee to improve or stop?
- Interpretation:
  - Results make sense? What groups? What principal component? Selected feature meaningful? What errors made by classifier? Improvements?

# Key mathematical tools

- Linear algebra and vector spaces
  - Vector, projection, linear combination
  - inner product, distance
  - Eigen-decomposition: $A = U\Sigma U^\top, or\ Av = \lambda v$
  - Singular value decomposition: $A = USV^\top, or\ Av = \sigma u$

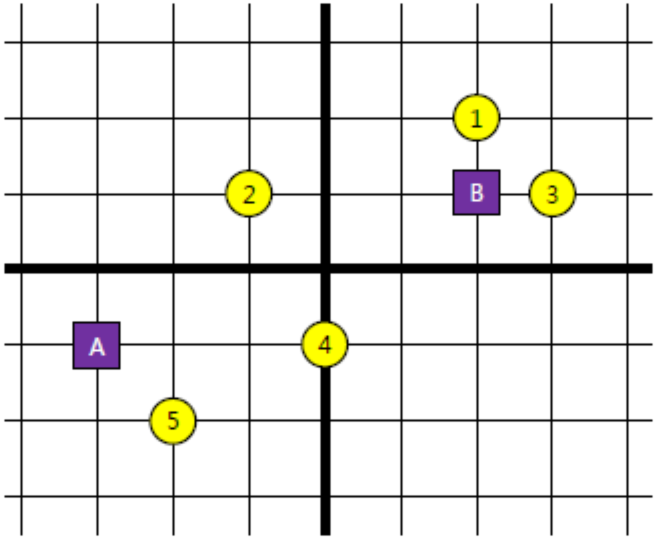
- Statistics
  - Mean, variance
  - Density, distribution, parametric models
  - Sum rule, product rule, Bayes rule
  - Maximum likelihood estimation
    - Fully observed case (often convex)
    - With hidden variables (expectation-maximization algorithm)

# Key mathematical tools (cont.)

- Optimization
  - Convex/concave function
  - Derivative of function (and with respect to vectors, matrices)
  - Lagrangian function
  - Optimality conditions
- Computer Science
  - Complexity: computation and memory, trade-off
  - Data structures: image and graph representation
  - Local search heuristic (greedy algorithms)
  - Sophisticated algorithm: shortest path, nearest neighbor search
  - Programming: loop vs. vectorized, underflow

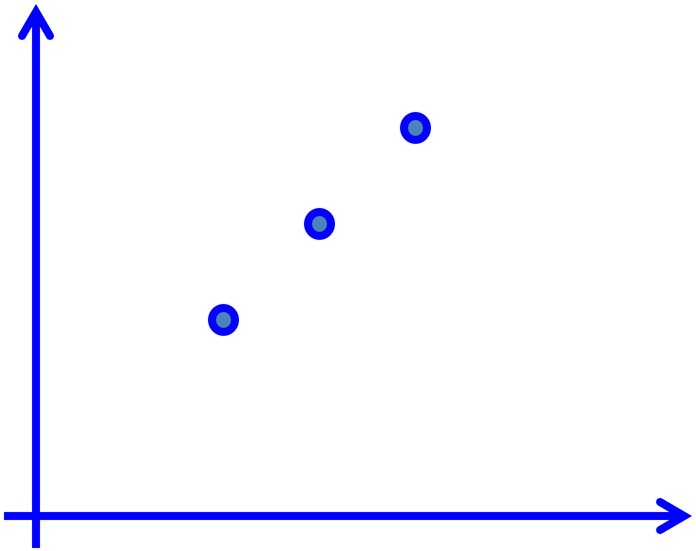
# Example question

- Use Euclidian distance, run one step:
  - Cluster assignment
  - New Center
- Will it terminate in one step?
- What about other distance?



# Example question

- Given you a few point
- What is the first principal axis?
- How about the second one?
- Represent the data using leading principal axis?
- What is the residue?



# Example question

- Given you a table
- How to estimate the parameter for $X_1$?
- How to estimate the joint probability of $X_1$ and $X_2$ with missing values?

| Example | $X_1$ | $X_2$ |
|---|---|---|
| 1 | 0 | 1 |
| 2 | 1 | 0 |
| 3 | 1 | 0 |
| 4 | 1 | ? |
| 5 | 0 | 1 |

sum log P(x1, x2) + log sum_x2 P(x1, x2)