Teoria dos Grafos
Aula 18

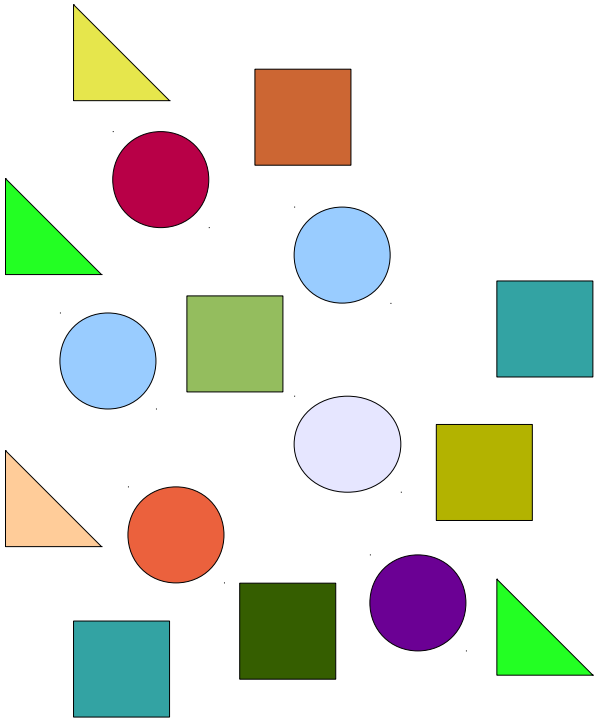
Aula passada

- Coloração
- Algoritmo guloso
- Número cromático
- Teorema das 4 cores

Aula de hoje

- Clusterização (ou agrupamento)
- Algoritmo
- Variação

Clusterização



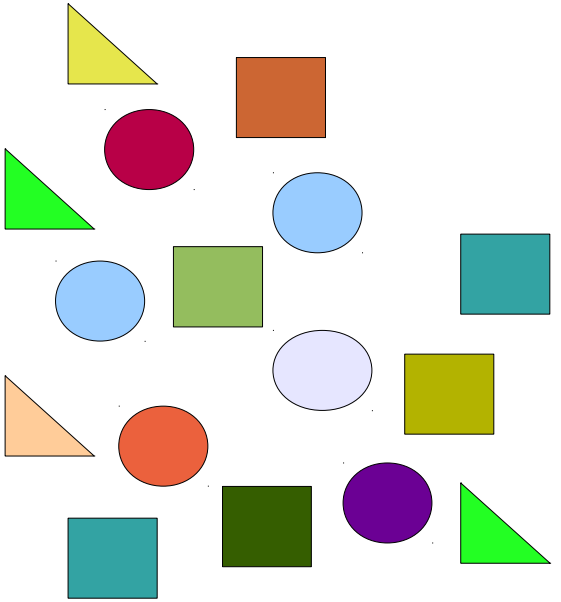
- Coleção de objetos
- Agrupar os objetos em conjuntos
- Objetos *similares* no mesmo conjunto

- **Problema:** Como agrupar os objetos da melhor maneira possível?

Problema fundamental (muitas aplicações)

Exemplos

- Fotografias
  - agrupamento de fotografias
- Biologia
  - agrupamento de espécies
- Web
  - agrupamento de páginas
- Redes sociais
  - agrupamento em comunidades
- Estrelas
  - agrupamento em galáxias
- Etc...



Medindo Similaridade

- **Problema**: como medir similaridade entre objetos?
- Definir uma função de distância entre os objetos
  - valor da função é inversamente proporcional a similaridade (menor valor, mais similar)
- $d(o_i, o_j)$: distância entre objetos o_i e o_j
  - $d(o_i, o_j) > 0$ se i != j
  - $d(o_i, o_i) = 0$
  - $d(o_i, o_j) = d(o_j, o_i)$ (simétrica)

Medindo Similaridade

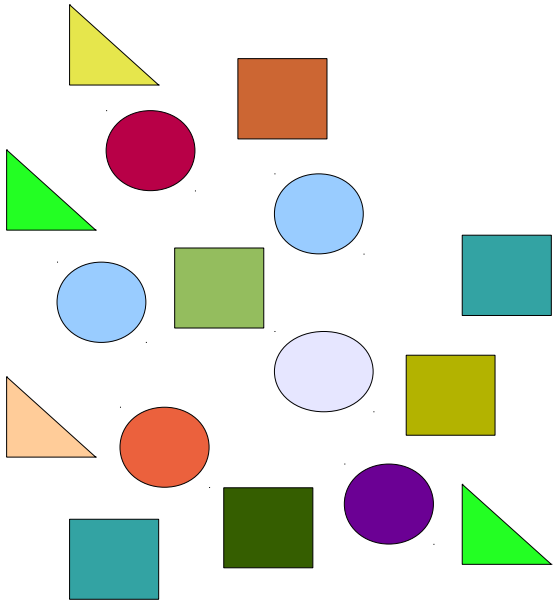
- Função de distância depende do domínio
- Distância entre fotografias
  - ex. número de pixels cuja diferença de cor é maior do que D
- Distância entre espécies
  - ex. percentagem de gens do DNA que são diferentes
- Distância entre documentos
  - ex. fração de palavras que são diferentes
- Etc.

Medindo a Clusterização

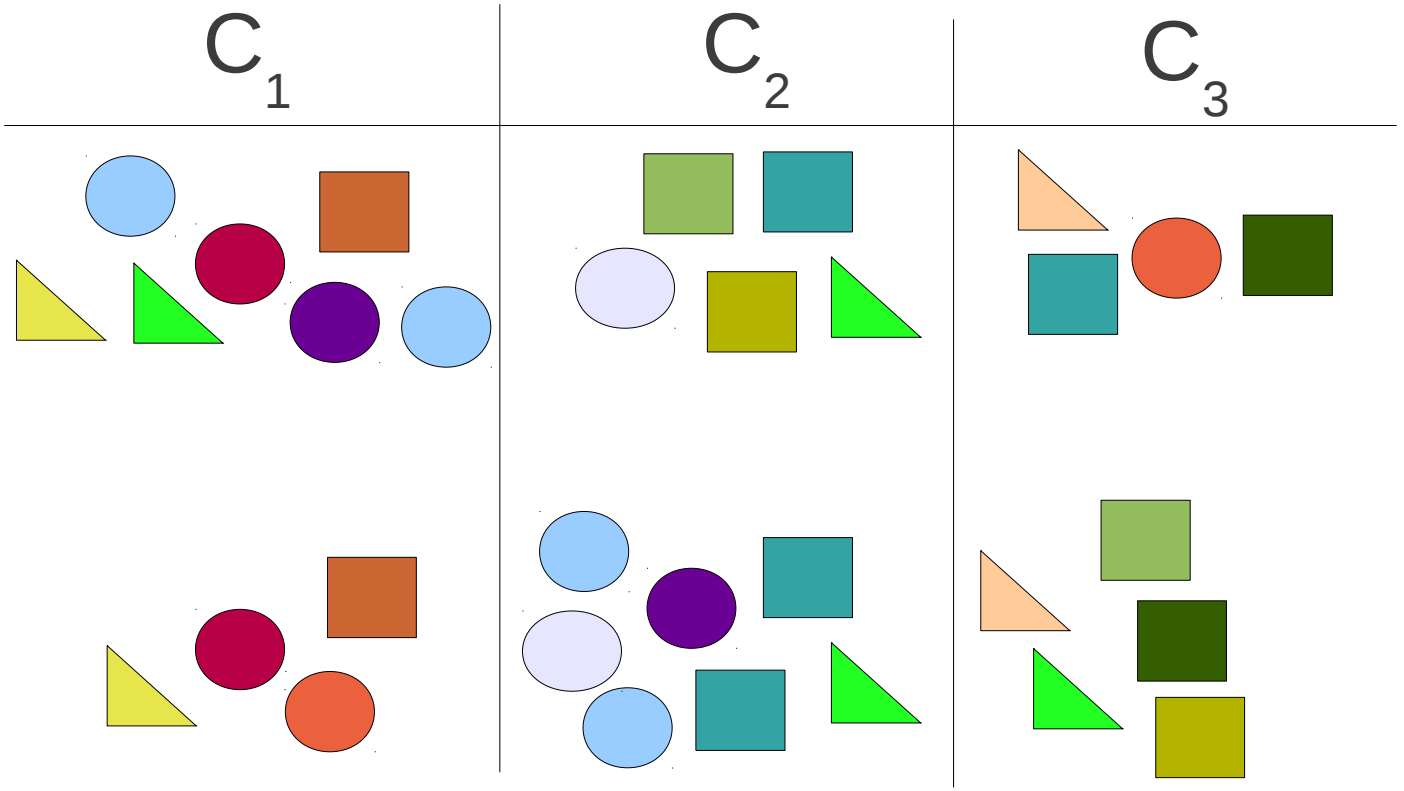
- K: número de conjuntos a serem produzidos
- **Problema**: como agrupar n objetos em K conjuntos?
- Existem muitas maneiras de agrupar n objetos em K conjuntos
  - número exponencial
- Qual delas é a melhor maneira?

Definir uma métrica para a qualidade da clusterização

Exemplo



- 16 objetos
- K = 3 (dividir em 3 conjuntos)



- Qual é a melhor clusterização?
- Depende!

Espaçamento de uma Clusterização

- Dado K conjuntos e as alocações
  - C_1, C_2, ..., C_K
- **Espaçamento**: menor distância entre quaisquer dois objetos de grupos diferentes
- Melhor agrupamento é aquele que tem maior espaçamento
- **Objetivo**: produzir um agrupamento ótimo (maximizar o espaçamento)

Abstração e Algoritmo

- Abstração via grafos
  - Vértices: objetos
  - Pesos nas arestas: distância entre objetos
- Grafo completo com n vértices (objetos)

Idéias para algoritmo?

- **Guloso**: agrupar objetos mais próximos primeiro
- Grupos são as componentes conexas

Algoritmo

- Começar com grafo totalmente desconexo
- Adicionar arestas em ordem crescente de peso
  - mais próximo primeiro
- Parar quando tivermos exatamente k componentes conexos
- Cada passo (cada aresta)
  - ou uni dois componentes conexos
  - ou adiciona aresta dentro de um componente conexo

Conexão com MST

- Similaridade com o MST?

Algoritmo de Kruskal

- Mesmo algoritmo, mas paramos antes
- Antes de adicionar as últimas k-1 arestas que Kruskal adicionaria
- Algoritmo equivalente: obter a MST e remover as K-1 mais pesadas

MST to the rescue!

Análise do Algoritmo

- Complexidade?
- Obter a MST (em um grafo completo) + remover K-1 arestas
  - Complexidade da MST

- Algoritmo produz agrupamento ótimo
- Componentes conexos obtidos pela remoção das K-1 arestas da MST constituem uma clusterização com K grupos com espaçamento máximo

Outra Métrica de Qualidade

- Dado K conjuntos e as alocações
  - C_1, C_2, ..., C_K
- Muitas métricas para definir a qualidade da clusterização
- Outra métrica: maior distância entre dois objetos de um mesmo grupo
  - Mede “diâmetro” de cada grupo
- **Objetivo:** produzir agrupamento ótimo (minimizar a métrica)

Dificuldade da Clusterização

- Métrica 1
  - Grupos cujos objetos diferentes estão distantes
  - Maximizar espaçamento entre grupos
- Métrica 2
  - Grupos cujos objetos estão próximos
  - Minimizar espaçamento intra grupos
- Métrica 1: algortimo polinomial (via MST)
- Métrica 2: não se conhece algoritmo polinomial (**surpreendente!**)
- Muitas heurísticas (ex. Algoritmo k-means)