Announcements
Review: MDL

- Learning as compression: find the way to compress data most efficiently.

- Minimum Description length (MDL) principle:
  - Estimate parameters $\hat{\theta}_M$ (ML, MAP etc.)
  - The two-stage MDL criterion for model selection:
    $$M^* = \arg\min_M DL\left(X_N | \hat{\theta}_M\right) + DL\left(\hat{\theta}_M\right).$$

- BIC is an asymptotic approximation for MDL, with $\hat{\theta}_M$ estimated/transmitted with precision $1/\sqrt{N}$.
Review: information theory

- Entropy of a discrete RV $X$ with multinomial pmf $p$, where $p_i = p(X = i)$:

$$H(p) = - \sum_{i=1}^{M} p_i \log p_i.$$

- Optimal code for $X$ achieves, asymptotically, $H(p)$ bits per symbol.

- Cost of coding $X$ with code optimal for $\hat{p} = q$ when the true pmf is $p$: the Kullback-Leibler divergence

$$D_{KL}(p\|q) \triangleq \sum_{i=1}^{m} p_i \log \frac{p_i}{q_i}.$$
Plan for today

- A few properties of KL-divergence
- Proof of EM monotonically increasing likelihood.
- Unsupervised learning: clustering
  - $k$-means
Properties of KL-divergence

- $D_{KL}(p\|q) \geq 0$ for any $p, q$
- $D_{KL}(p\|q) = 0$ if and only if $p \equiv q$
- It’s asymmetric:
  - If $p_i = 0$, $q_i \geq 0$
Properties of KL-divergence

• \( D_{KL}(p\|q) \geq 0 \) for any \( p, q \)

• \( D_{KL}(p\|q) = 0 \) if and only if \( p \equiv q \)

• It’s asymmetric:
  – If \( p_i = 0, q_i \geq 0 \Rightarrow 0 \cdot \log(0) \rightarrow 0. \)
Properties of KL-divergence

• $D_{KL}(p∥q) \geq 0$ for any $p, q$

• $D_{KL}(p∥q) = 0$ if and only if $p ≡ q$

• It’s asymmetric:
  – If $p_i = 0, q_i \geq 0 \Rightarrow 0 \cdot \log(0) \to 0$
  – If $q_i = 0, p_i \geq 0$
Properties of KL-divergence

- $D_{KL}(p\|q) \geq 0$ for any $p, q$
- $D_{KL}(p\|q) = 0$ if and only if $p \equiv q$
- It’s asymmetric:
  - If $p_i = 0$, $q_i \geq 0 \Rightarrow 0 \cdot \log(0) \rightarrow 0$.
  - If $q_i = 0$, $p_i \geq 0 \Rightarrow p_i \cdot \log(p_i/0) \rightarrow \infty$. 

Properties of KL-divergence

- $D_{KL}(p\|q) \geq 0$ for any $p, q$
- $D_{KL}(p\|q) = 0$ if and only if $p \equiv q$
- It’s asymmetric:
  - If $p_i = 0$, $q_i \geq 0$ $\Rightarrow 0 \cdot \log(0) \rightarrow 0$.
  - If $q_i = 0$, $p_i \geq 0$ $\Rightarrow p_i \cdot \log(p_i/0) \rightarrow \infty$.
- Continuous KL-divergence:

$$D_{KL}(p\|q) = \int p(x) \log \frac{p(x)}{q(x)} dx.$$
Back to EM

- **Recall**: $X$ are observed, $Z$ are hidden; by chain rule

\[
p(X, Z \mid \theta) = p(Z \mid X, \theta) p(X \mid \theta)
\]

\[
\log p(X, Z \mid \theta) - \log p(Z \mid X, \theta) = \log p(X \mid \theta)
\]

- **Now take expectation w.r.t. $p(Z \mid X, \theta^{old})$**:

\[
E_{p(Z \mid X, \theta^{old})} [\log p(X, Z \mid \theta)] - E_{p(Z \mid X, \theta^{old})} [\log p(Z \mid X, \theta)] = \log p(X \mid \theta)
\]
Likelihood of EM solution

\[ \log p (X \mid \theta) = Q(\theta; \theta^{old}) \quad E_p(Z \mid X, \theta^{old}) [\log p (Z \mid X, \theta)] \]

- Since \( \theta^{new} = \text{argmax}_\theta Q(\theta; \theta^{old}) \), we have \( Q(\theta^{new}; \theta^{old}) \geq Q(\theta^{old}; \theta^{old}) \).

- Also,

\[
E_p(Z \mid X, \theta^{old}) [\log p (Z \mid X, \theta^{old})] \quad E_p(Z \mid X, \theta^{old}) [\log p (Z \mid X, \theta^{new})] \\
= \int p(Z \mid X, \theta^{old}) \log \frac{p(Z \mid X, \theta^{old})}{p(Z \mid X, \theta^{new})} \\
= D_{KL} (p(Z \mid X, \theta^{old}) \mid \mid p(Z \mid X, \theta^{new})) \geq 0.
\]

- So,

\[ p(X \mid \theta^{new}) - p(X \mid \theta^{old}) \geq 0. \]
Unsupervised learning

- Three main types of problems in unsupervised learning:
  - Density estimation: learning a density function from a few samples.
    - Closed-form ML or MAP estimation for Gaussian, Bernoulli models; EM for mixture models.
  - Clustering: grouping similar training cases together.
  - Dimensionality reduction: learning to represent each training case using a small number of continuous variables from which the original data can be almost exactly reconstructed.

- For a clustering problem to be well-defined, we need to decide what “similar” means.
• We have discussed mixture models as models for density estimation.
  – The goal has been to estimate $p(x|\theta)$; the hidden variables $z$ were for convenience.

• What if we only care about $z$, i.e. which component generated which $x$?
  – Clustering problem: assign each example to a group.
Gaussian mixture in the limit

- Suppose we fix $\Sigma_c = \sigma^2 I$, for all $c = 1, \ldots, k$ and set $\sigma^2 \to 0$.

- Suppose $c^* = \arg\min_c \|x_i - \mu_c\|$ (the closest mean).

- The responsibility become “winner take all”.

$$\lim_{\sigma^2 \to 0} \gamma_{ic} = \begin{cases} 0 & \text{if } c \neq c^*; \\ 1 & \text{if } c = c^*. \end{cases}$$
$k$-means clustering

1. Initialize $k$ means $\mu_1, \ldots, \mu_k$ to random locations.
   - E.g., set to $k$ randomly chosen distinct examples.

2. Repeat until no change in assignment:
   - **E-step:** Assign each example to the closest mean:
     \[ y_i = \arg\min_c \| x_i - \mu_c \|. \]
   - **M-step:** Reestimate each mean based only on examples assigned to it:
     \[ N_c = |\{ x_i : y_i = c \}| ; \quad \mu_c = \frac{1}{N_c} \sum_{y_i = c} x_i. \]
$k$-means: example
$k$-means: example
$k$-means: example
$k$-means: example
$k$-means: example
$k$-means as optimization

- What objective is optimized in $k$-means?
What objective is optimized in $k$-means?

The “ideal” objective: minimum squared

$$J^*(\mu) = \frac{1}{N} \sum_{i=1}^{N} \min_{c=1,\ldots,k} (x_i - \mu_c)^T(x_i - \mu_c).$$
$k$-means as optimization

- What objective is optimized in $k$-means?

- The “ideal” objective: minimum squared

\[
J^*(\mu) = \frac{1}{N} \sum_{i=1}^{N} \min_{c=1, \ldots, k} (x_i - \mu_c)^T (x_i - \mu_c).
\]

- That’s intractable; instead, $k$-means optimizes the upper bound

\[
J(\mu, \{y_i\}) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_{y_i})^T (x_i - \mu_{y_i}).
\]
Vector Quantization

- We can use the cluster mean as a prototype representing all the examples assigned to the cluster.

- **Vector quantization**: construct a codebook using $k$-means.

- Whenever need to transmit $x$, transmit instead the closest codebook.
  - The bits to transmit: $\log kd$ once + $\log k$ for every message.
Setting $k$

- How can we set $k$?
Setting $k$

- How can we set $k$? Cross-validation doesn’t work (why?)
Setting $k$

- How can we set $k$? Cross-validation doesn’t work (why?)

- The relevant statistic: *within-class dissimilarity*

$$W_k = \sum_{c=1}^{k} \sum_{y_i = y_j = c} \|x_i - x_j\|^2.$$  

- A popular (heuristic) strategy: look for an “elbow” in $W_k$
Setting $k$

- How can we set $k$? Cross-validation doesn’t work (why?)

- The relevant statistic: *within-class dissimilarity*

  $$W_k = \sum_{c=1}^{k} \sum_{y_i=y_j=c} \|x_i - x_j\|^2.$$  

- A popular (heuristic) strategy: look for an “elbow” in $W_k$
Mixture of Gaussians EM versus $k$-means

- $k$-means:
  - No probabilistic model $\Rightarrow$ no estimated density.
  - Faster to compute (only a single explanation for each data point).
  - Limited by the underlying assumption of spherical clusters.

- We can bring back the covariance—get “hard EM”.
  - Still limited by the shape of the covariance (ellipsoid).

- Both EM and $k$-means depend on initialization (can get stuck in local optima).
Next time

Other clustering methods.