Generative Models and EM

Harrison Wang
11/4/22
MCB 112
Outline

Generative models

EM

\[ \hat{c}_k = \sum_{n=1}^{N} q_{nk} = \sum_{n=1}^{N} q_{nk}(\hat{v}_k) \]

\[ \hat{v}_k = \frac{\hat{c}_k}{N} \]
Outline

Generative models

\[ \hat{c}_k = \sum_{n=1}^{N} q_{nk} = \sum_{n=1}^{N} q_{nk}(\hat{\nu}_k) \]

\[ \hat{\nu}_k = \frac{\hat{c}_k}{N} \]
Generative models

• variable and parameter types
• How each are generated

We will go through these with examples.
example: negative binomial fitting

data

generative model
Generative models: parameter types

• Unknown parameters
• Known parameters
• Hidden variables
• Observed data
parameter types: negative binomial fitting

- Unknown parameters: centroids and mixture coefficients $\mu_k, \pi_k$
- Known parameters: dispersion $\phi$
- Hidden variables: group identity $G_n$
- Observed data: data points $x_n (1 \leq n \leq N)$
parameter types: RNA-seq experiment

- Unknown parameters: nucleotide abundances $\nu_1, \nu_2, \cdots, \nu_M$
- Known parameters: transcript lengths $L_1, L_2, \cdots, L_M$
- Hidden variables: identity, orientation, start $G_n, S_n, O_n$
- Observed data: reads $R_n (1 \leq n \leq N)$
Generative models: specifications

• Which parameters are known and unknown?

• How do the parameters generate the hidden variables?

• How do the parameters and hidden variables generate the observations?
Specifications: example (mixture NB)

- Which parameters are known and unknown?

\[ \mu_k, \pi_k, \phi \]

- How do the parameters generate the hidden variables \( G_n \)?

\[ P(G_n = j) = \pi_j \]

- How do the parameters and hidden variables generate the observations?

\[ P(x_n|G_n = j) \sim \mathcal{NB}(\mu_j, \sigma^2) \]
Outline

Generative models

EM

\[ \hat{c}_k = \sum_{n=1}^{N} q_{nk} = \sum_{n=1}^{N} q_{nk}(\hat{v}_k) \]

\[ \hat{v}_k = \frac{\hat{c}_k}{N} \]
EM

• Initialization
  • Make any guess for the unknown parameters.

• Expectation
  • use the generative model!

• Maximization
  • Find the unknown parameter values that maximize the likelihood.
EM: negative binomial fitting

• Initialization
  • Make any guess for the unknown parameters $\mu_k, \pi_k$.

• Expectation
  • use the generative model!
    \[ P(G_n = j) = \pi_j \]
    \[ P(x_n|G_n = j) \sim \mathcal{NB}(\mu_j, \sigma^2) \]
  • Then find the posterior.
    \[ q_{nk} = P(G_n = k|x_n) = \frac{P(x_n|G_n = k)P(G_n = k)}{\sum_{j=1}^{K} P(x_n|G_n = j)P(G_n = j)} \]
EM: negative binomial fitting

• Maximization
  • Find the values of the unknown parameters that maximize the likelihood.
  • In pset 5: rather than writing out the likelihood and then maximizing it, we took a point estimate by weighting with the posterior.

\[
\hat{\mu}_k = \frac{\sum_{n=1}^{N} x_n \cdot q_{nk}}{\sum_{n=1}^{N} q_{nk}}
\]

\[
\hat{\pi}_k = \frac{1}{N} \sum_{n=1}^{N} q_{nk}
\]

• Now we have point estimates for the unknown parameters that we can plug back into the expectation step.
EM: RNA-seq

- Initialization
  - Make a sensible guess for the unknown parameters $\nu_1, \nu_2, \cdots, \nu_M$.
  - Now you have a complete set of parameters $\theta$.

- Expectation
  - use the generative model!
  - Then find the posterior.
    \[
    q_{nk} = P(G_n = k | R_n, \theta) = \frac{P(R_n | G_n = k, \theta)P(G_n = k | \theta)}{\sum_{j=1}^{K} P(R_n | G_n = j, \theta)P(G_n = j | \theta)}
    \]

What is the probability of generating a transcript with a certain identity?
What is the probability of generating this read, given the transcript identity?
EM: RNA-seq

• Maximization
  • Find the values of the unknown parameters that maximize the likelihood.
  • Again, we take a point estimate by "weighting" with the posterior.

\[
\hat{c}_k = \sum_{n=1}^{N} q_{nk} = \sum_{n=1}^{N} q_{nk}(\hat{\nu}_k)
\]

\[
\hat{\nu}_k = \frac{\hat{c}_k}{N}
\]

• Now we have point estimates for the unknown parameters that we can plug back into the expectation step.
Resources for this week

• In these slides, there’s a sketch of how to do pset part 3.
• You can refer to past resources to see how EM was implemented for negative binomial fitting.
• Section notes contains a sketch of how to write an array for calculating the expectation.