Announcements
Review: AdaBoost

1. Initialize weights: \( W_i^{(0)} = 1/N \)

2. Iterate for \( m = 1, \ldots, M \):
   - Find (any) “weak” classifier \( h_m \) that attains weighted error
     \[
     \epsilon_m = \frac{1}{2} \left( 1 - \sum_{i=1}^{N} W_i^{(m-1)} y_i h_m(x_i) \right) < \frac{1}{2}
     \]
   - Let \( \alpha_m = \frac{1}{2} \log \frac{1-\epsilon_m}{\epsilon_m} \).
   - Update the weights and normalize so that \( \sum_i W_i^{(m)} = 1 \):
     \[
     W_i^{(m)} = \frac{1}{Z} W_i^{(m-1)} e^{-\alpha_m y_i h_m(x_i)},
     \]

3. The combined classifier: \( \text{sign} \left( \sum_{m=1}^{M} \alpha_m h_m(x) \right) \)
Plan for today

- Mixtures of experts
- Sequential data
  - Markov models
Mixture model for regression

- Example:
Mixture model for regression

- Example:

- We can represent this as a mixture of two regression models
  - Two experts;
  - Need to switch between them according to $x$. 
Mixture of experts model

- Expert $j$ holds a parameteric model $p(y \mid x; \theta_j)$, e.g.,

  $\theta_j = \{w_j, \sigma_j^2\}$,

  $p(y \mid x; \theta_j) = \mathcal{N}(y; w_j^T x, \sigma_j^2)$

- The distribution of $y$ is a *conditional mixture* model:

  $p(y \mid x; \theta) = \sum_{j=1}^{c} p(j \mid x) p(y \mid x; \theta_j)$. 
Gating network

\[ p(y | x; \theta) = \sum_{j=1}^{c} p(j | x) p(y | x; \theta_j) \]

- A gating network specifies the conditional distribution \( p(j | x; \eta) \)

- Possible gating functions:
  - \( k = 2 \) experts: logistic regression, parametrized by \( \eta = \{v\} \)
    \[ p(j | x; \eta) = \left( 1 + e^{-v^T x} \right)^{-1} \]
  - \( k > 2 \) experts: softmax, \( \eta = \{v_1, \ldots, v_k\} \)
    \[ p(j | x; \eta) = \frac{e^{v_j^T x}}{\sum_{t=1}^{k} e^{v_t^T x}}. \]
Conditional mixtures

- **Parametrization**
  - Regression models $p(y \mid x; \theta_j)$
    e.g., linear regressors, $\theta_j = \{w_j, \sigma_j^2\}$.
  - Gating network $p(j \mid x; \eta)$
    e.g., logistic regression, $\eta = \{v\}$
Learning a MoE model

- The graphical model:

- Responsibilities:

\[
\gamma_{ij} = p(j \mid x_i, y_i; \theta, \eta) = \frac{p(j \mid x_i; \eta)p(y_i \mid x_i; \theta_j)}{\sum_{c=1}^{k} p(c \mid x_i; \eta)p(y_i \mid x_i; \theta_c)}
\]
EM for mixtures of experts

**Initialize** random $\theta_j, \sigma_j^2, \eta$.

**E-step** Compute responsibilities $\gamma_{ij} = p(j \mid x_i, y_i; \theta^{old}, \eta^{old})$

**M-step** Separately:

- For each expert $j$ estimate
  $$\theta_j^{new} = \arg\max_{\theta} \sum_{i=1}^{N} \gamma_{ij} \log p(y_i \mid x_i; \theta)$$

- Estimate the gating network:
  $$\eta^{new} = \arg\max_{\eta} \sum_{i=1}^{N} \sum_{j=1}^{k} \gamma_{ij} \log p(j \mid x_i; \eta)$$
EM for mixtures of experts: example

Iter 1

Iter 2
EM for mixtures of experts: example

Iter 3

Iter 7
More on mixtures of experts

- MoE for classification: similar idea.
  - Compare to AdaBoost; what are the similarities/differences?

- Hierarchical MoE: multiple levels of gating.
Sequential data

• Departure from the i.i.d. assumption:
  – Probability of observing \( x_i \) depends on \( x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N \).

• The sequential dimension may be temporal or spatial:
  – Speech (measurements of acoustic waveform);
  – Language (words);
  – Images (pixels).

• Almost always: assume dependence on past only.

\[
p(x_i | x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N) = p(x_i | x_1, \ldots, x_{i-1}).
\]

• Still, complexity grows as we increase \( N \).
Markov models

- The $k$-th order Markov model:

$$p(x_i | x_1, \ldots, x_{i-1}) = p(x_i | x_{i-k}, \ldots, x_{i-1}).$$

- Zeroth order:
Markov models

- The $k$-th order Markov model:

$$p(x_i | x_1, \ldots, x_{i-1}) = p(x_i | x_{i-k}, \ldots, x_{i-1}).$$

- Zeroth order:

- First order (bigrams):
Markov models

• The $k$-th order Markov model:

$$p(x_i | x_1, \ldots, x_{i-1}) = p(x_i | x_{i-k}, \ldots, x_{i-1}).$$

• Zeroth order:

• First order (bigrams):

• Second order (trigrams):
Dynamic models

• Suppose \( X = x_1, \ldots, x_N \) generated by 1-st order Markov process.

\[
p(X) = p(x_1)p(x_2 | x_1)p(x_3 | x_2) \cdots p(x_N | x_{N-1}).
\]

• If \( p(x_{t+1} | x_t) \) does not depend on \( i \), the Markov model is \textit{homogenous}.

• \textit{Discrete} observations called \textit{states} \( s_t \in \{1, \ldots, m\} \), model is parametrized by:
  – Starting probability \( s_1 \sim p_0 \): a \( m \times 1 \) vector.
  – Transition probability matrix \( P \):

\[
P_{ij} = p(s_{t+1} = j | s_t = i).
\]
Representing discrete Markov models

Three equivalent graphical representations:

- Graphical model
Representing discrete Markov models

Three equivalent graphical representations:

- **Graphical model**

- **State transition diagram**
Representing discrete Markov models

Three equivalent graphical representations:

- **Graphical model**

- **State transition diagram**

- **Trellis**
Markov model: example

\[
p_0 = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}, \quad P = \begin{bmatrix} 0.5 & 0.5 \\ 0 & 1 \end{bmatrix}
\]

- Reminder: \( p(s_{t+k} = i \mid s_t = j) = (P^k)_{ij} \)

- Reminder: Markov chain is \textit{ergodic} if \((P^k)_{ij} > 0\) for all \(i, j\) and some fixed \(k\).
  - Cf. random walk on affinity graph in spectral clustering.
Estimating Markov model parameters

- Need to estimate \( p_0, \mathbf{P} \).

- Log-likelihood of observed \( s_1, \ldots, s_N \): 
  \[
  \log p_0(s_1) + \sum_{t=2}^{N} \log p(s_t | s_{t-1})
  \]

- ML for \( \mathbf{P} \): let \( n_{r \rightarrow s} \) be the \# of times \( s_{t-1} = r, s_t = s \).

  \[
  \hat{P}_{ij} = \frac{n_{i \rightarrow j}}{\sum_r n_{i \rightarrow r}}
  \]

- ML for \( p_0 \): trivial if we have \( L > 1 \) sequences.

  \[
  \hat{p}_0(s) = (\# \text{ of times } s_1 = s)/L.
  \]
Estimating Markov model parameters

- What if we only have a single sequence?
Estimating Markov model parameters

- What if we only have a single sequence?

\[ \hat{p}_0(s) = \frac{1}{N} \sum_i n_{i \rightarrow s} \]
Estimating Markov model parameters

- What if we only have a single sequence?

\[ \hat{p}_0(s) = \frac{1}{N} \sum_i n_{i \rightarrow s} \]

- Is ML a good estimator for \( P \)?
  - With \( m \) states, \( P \) has \( m^2 \) parameters.
Markov models for language

- $k$-th order Markov model is also called a $k$-gram model

- Example (C. Shannon): character $k$-grams as a generative model.

  $k = 0$  
  XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGXYD QPAAMKBZAACIBZLHJQD
Markov models for language

- $k$-th order Markov model is also called a $k$-gram model

- Example (C. Shannon): character $k$-grams as a generative model.

  $k = 0$   XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGXYD
            QPAAMKBZAACIBZLHJQD

  $k = 1$   OCRO HLI RGWR NMIELWIS EU LL NBBESEBYA TH EEI ALHENHTTPA
            OO BTTV
Markov models for language

- $k$-th order Markov model is also called a $k$-gram model

- Example (C. Shannon): character $k$-grams as a generative model.

  $k = 0$  XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGXYD
            QPAAMKBZAACIBZLHJQD
  $k = 1$  OCRO HLI RGWR NMIELWIS EU LL NBBESEBYA TH EEI ALHENHTTPA
            OO BTTV
  $k = 2$  ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN
            D ILONASIVE TUCOOWE FUSO TIZIN ANDY TOBE SEACE CTISBE
Markov models for language

- $k$-th order Markov model is also called a $k$-gram model

- Example (C. Shannon): character $k$-grams as a generative model.

  \[
  k = 0 \quad \text{XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGXYD}
  \quad \text{QPAAMKBZAACIBZLHJQD}
  \]

  \[
  k = 1 \quad \text{OCRO HLI RGWR NMIELWIS EU LL NBBESEBYA TH EEI ALHENHTTPAOO BTTV}
  \]

  \[
  k = 2 \quad \text{ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHINDILONASIVE TUOCOWEFUSOTIZINANDYTOBESACECTISBE}
  \]

  \[
  k = 3 \quad \text{IN NO IST LAY WHEY CRATICT FROURE BERS GROCID PONDENOMEOF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE}
  \]
Markov models for language

- $k$-th order Markov model is also called a $k$-gram model

- Example (C. Shannon): character $k$-grams as a generative model.

$k = 0$  
XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGXYD
  QPAAMKBZACIBZLHJQD

$k = 1$  
OCRO HLI RGWR NMIELWIS EU LL NBBESEBYA TH EEI ALHENHTTPA
  00 BTTV

$k = 2$  
ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN
  D IلونAIVE TUOOWE FUSO TIZIN ANDY TOBE SEACE CTISBE

$k = 3$  
IN NO IST LAY WHEY CRATICT FROURE BERS GROCID PONDENOME
  OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE

$k = 4$  
THE GENERATED JOB PROVIDUAL BETTER TRAND THE DISPLAYED
  CODE ABOVERY UPONDULTS WELL THE CODERST IN THESTICAL IT TO
  HOCK BOTHE
Next time

Hidden Markov Models