Acknowledgement:
Many of these slides are derived from Tom Mitchell, Pascal Poupart, Pieter Abbeel, Eric Eaton, Carlos Guestrin, William Cohen, and Andrew Moore.
Training logistic regression: MCLE

- We need to choose $W = \langle w_0, \ldots, w_n \rangle$ to maximize the conditional likelihood of training data

  where $P(Y = 0 \mid X, W) = \frac{1}{1 + \exp(w_0 + \sum_{i=1}^{n} w_i x_i)}$

  and $P(Y = 1 \mid X, W) = \frac{\exp(w_0 + \sum_{i=1}^{n} w_i x_i)}{1 + \exp(w_0 + \sum_{i=1}^{n} w_i x_i)}$

- Training data $D = \{ \langle X^1, Y^1 \rangle, \ldots, \langle X^L, Y^L \rangle \}$

  Data likelihood is $\prod_{l} P( \langle X^l, Y^l \rangle \mid W)$

  Data conditional likelihood is $\prod_{l} P(Y^l \mid X^l, W)$

  Therefore we need to estimate $W_{MCLE} = \arg \max_{W} \prod_{l} P(Y^l \mid X^l, W)$
Expressing conditional log likelihood

\[ l(W) = \ln \prod_l P(Y^l | X^l, W) = \sum_l \ln P(Y^l | X^l, W) \]

where \[ P(Y = 0 | X, W) = \frac{1}{1 + \exp(w_0 + \sum_{i=1}^n w_i x_i)} \]

and \[ P(Y = 1 | X, W) = \frac{\exp(w_0 + \sum_{i=1}^n w_i x_i)}{1 + \exp(w_0 + \sum_{i=1}^n w_i x_i)} \]

\[ l(W) = \sum_l Y^l \ln P(Y^l = 1 | X^l, W) + (1 - Y^l) \ln P(Y^l = 0 | X^l, W) \]

\[ = \sum_l Y^l \ln \frac{P(Y^l = 1 | X^l, W)}{P(Y^l = 0 | X^l, W)} + \ln P(Y^l = 0 | X^l, W) \]

\[ = \sum_l Y^l (w_0 + \sum_{i=1}^n w_i x_i^l) - \ln(1 + \exp(w_0 + \sum_{i=1}^n w_i x_i^l)) \]
Maximizing conditional log likelihood

\[ l(W) = \ln \prod_{l} P(Y^l | X^l, W) \]

\[ = \sum_{l} Y^l(w_0 + \sum_{i} w_iX^l_i) - \ln(1 + \exp(w_0 + \sum_{i} w_iX^l_i)) \]

- **Good news:** \( l(W) \) is a concave function of \( W \)
- **Bad news:** no closed-form solution to maximize \( l(W) \)
Gradient ascent

\( E(\vec{W}) \)

\( \vec{w}_1 \)

\( \vec{w}_2 \)

\( \nabla E(\vec{w}) = \left[ \frac{\partial E}{\partial w_0}, \frac{\partial E}{\partial w_1}, \ldots, \frac{\partial E}{\partial w_n} \right] \)

Training Rule: \( \vec{w}^{(i+1)} \leftarrow \vec{w}^i + \eta \nabla E(\vec{w}) \)

i.e. \( \Delta w_i = \eta \frac{\partial E}{\partial w_i} \)
Maximizing conditional log likelihood via gradient ascent

\[ l(W) = \ln \prod_l P(Y^l | X^l, W) \]
\[ = \sum_l Y^l(w_0 + \sum_i^n w_i X_i^l) - \ln(1 + \exp(w_0 + \sum_i^n w_i X_i^l)) \]

\[ \frac{\partial l(W)}{\partial w_i} = \sum_l X_i^l(Y^l - \hat{P}(Y^l = 1 | X^l, W)) \]

- **Gradient ascent algorithm**: iterate until change < \( \varepsilon \)
  - \( \forall i \) repeat \( w_i \leftarrow w_i + \eta \sum_l X_i^l(Y^l - \hat{P}(Y^l = 1 | X^l, W)) \)
  - assume \( X_0 = 1 \) for \( w_0 \) step size (a.k.a. learning rate)