CS 4824/ECE 4424: Regression

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

- So far, we’ve been interested in learning $P(Y|X)$ where $Y$ has discrete values (called ‘classification’)

- What if $Y$ is continuous? (called ‘regression’)
  - predict snow/rainfall from current and past weather features
  - predict stock price from current and past market conditions
  - predict weight from gender, height, age, ...
Regression: problem setting

- Wish to learn $f: X \rightarrow Y$ where $Y$ is real-valued, given training data
  $\{<X^1, Y^1> \ldots <X^n, Y^n>\}$

- Approach:
  - Choose some parameterized form for $P(Y|X, \theta)$ where $\theta$ is the vector of parameters
  - Estimate $\theta$ using MLE or MAP estimation
Choose parameterized form for $P(Y | X, \theta)$

- Assume $Y$ is some deterministic $f(X)$, plus random noise
  - $y = f(x) + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, \sigma)$

- Therefore, $Y$ is a random variable that follows the distribution
  - $p(y | x) = \mathcal{N}(f(x), \sigma)$

- And the expected value of $y$ for any given $x$ is $f(x)$
Consider linear regression

- \( p(y \mid x) = \mathcal{N}(f(x), \sigma) \)

- Assume \( f(x) \) is a linear function of \( x \), i.e.,
  \[ f(x) = \omega_0 + \omega_1 x \]

- \( p(y \mid x) = \mathcal{N}(\omega_0 + \omega_1 x, \sigma) \)

- \( \mathbb{E}(y \mid x) = \omega_0 + \omega_1 x \)
Consider linear regression

- \( p(y \mid x) = \mathcal{N}(f(x), \sigma) \)

- Assume \( f(x) \) is a linear function of \( x \) i.e., \( y = w_0 + w_1 x \)
  - \( p(y \mid x) = \mathcal{N}(w_0 + w_1 x, \sigma) \)
  - \( \mathbb{E}(y \mid x) = w_0 + w_1 x \)

- Note: to make our parameters explicit, let’s write
  - \( W = <w_0, w_1> \)
  - \( p(y \mid x, W) = \mathcal{N}(w_0 + w_1 x, \sigma) \)
Training linear regression

- $p(y | x) = \mathcal{N}(f(x), \sigma)$

- How can we learn $W$ from data?
Training linear regression

- $p(y \mid x) = \mathcal{N}(f(x), \sigma)$

- How can we learn $W$ from data?

- Learn $W$ using Maximum Conditional Likelihood Estimation!
  - $W_{MCLE} = \arg \max_W \prod_l P(Y^l \mid X^l, W)$
  - $W_{MCLE} = \arg \max_W \sum_l \ln p(y^l \mid x^l, W)$
  - Where $P(y \mid x, W) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{y^l - f(x, W)}{\sigma} \right)^2}$
**MCLE derivation**

\[
W_{MCLE} = \arg\max_W \sum_l \ln p(y^l | x^l, W)
\]

\[
P(y | x, W) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{y^l - f(x, W)}{\sigma} \right)^2}
\]

or

\[
P(y | x, W) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{y^l - (w_0 + w_1 x^l)}{\sigma} \right)^2}
\]

\[
W_{MCLE} = \arg\max_W \sum_l \left[ \ln \frac{1}{\sigma \sqrt{2\pi}} + \left( \frac{1}{2} \left( \frac{y^l - (w_0 + w_1 x^l)}{\sigma} \right)^2 \right) \right]
\]

\[
= \arg\max_W -\frac{1}{2\sigma^2} \left( y^l - (w_0 + w_1 x^l) \right)^2
\]

\[
W_{MCLE} = \arg\min_W \sum_l \left[ \frac{1}{2\sigma^2} \left( y^l - (w_0 + w_1 x^l) \right)^2 \right]
\]

\[
W_{MCLE} = \arg\min_W \left( y^l - (w_0 + w_1 x^l) \right)^2
\]

\[
W_{MCLE} = \text{MSE}^{\text{Squared error}}
\]
Training linear regression

- Learn $W$ using Maximum Conditional Likelihood Estimation:
  \[ W_{MCLE} = \arg \min_W \sum_l (y - f(x, W))^2 \]

- Training linear regression involves minimizing a loss function capturing the squared error (often used in “curve fitting”)