ECE/SIOC 228 Machine Learning for Physical Applications

Lecture 2: Supervised Learning Setup

Instructor: Yuanyuan Shi

Teaching Assistants:
Srinivas Rao Daru, sdaru@ucsd.edu
Tawaana Homavazir, thomavaz@ucsd.edu
Rohin Garg, rgarg@ucsd.edu
Rishabh Jangir rjangir@ucsd.edu,

Logistics

Exam:

- Midterm exam will be on May 17th Tuesday (2:00pm - 3:20pm)
- It will be an closed-book exam and you are allowed to bring a A4 paper written note (two-sided)
- Based on what is covered on lectures

Office hour update:

- My office hour: 4-5pm Thursdays
- Jacobs Hall Room 4401

Survey Results

Logistics

Project

- Many of you have already started to form teams - that is great!!
- For those haven't started - start early :)
- Feel free to discuss your idea at any of the TAs' office hour and my office hour
- Feel free to reach out to your assigned TA for idea discussion if you're not able to attend their office hours; we'll adjust the TA assignment after Week 3 when the teams are formed.
- Please don't share individual docs and writeup with us - the proposal is due on Week 3 and we won't be able to provide feedback on writings before then.

Today's Lecture

- Supervised Learning Setup
- Linear Regression
- Logistic Regression
- Computation Graph

Reference & Acknowledgement: Stanford CS230 Deep Learning Lecture Note, Andrew Ng

Supervised Learning Setup

Suppose we have m pairs of data points: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), ..., (x^{(m)}, y^{(m)})\}$, where for each data $(x^{(i)}, y^{(i)})$: Specifically,

- $x^{(i)} \in \mathbb{R}^n$ is the input vector for the i^{th} sample, and $\mathbb{R}^n$ is the n-dimension feature space
- $y^{(i)} \in C$ is the label of the i^{th} sample and C is the label space

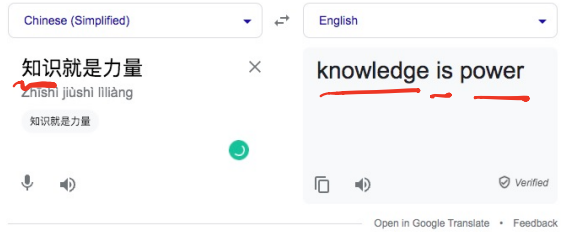
Example of label spaces:

- $C = \mathbb{R}$ Regression: predicting the electricity consumption of a house given number of rooms, square feet, outside temperature
- $C = \{0, 1\}$ Binary classification: predicting whether this email is a spam (1) or not (0).
- $C = \{1, 2, ..., K\}$ Multi-class classification: CIFAR-10 (cat, deer, dog, frog, horse, bird, airplane, ...)

Supervised Learning Setup

Examples of feature vectors:

- Structured data
  - Electricity consumption data. $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, ..., x_n^{(i)})$, where $x_1^{(i)} = \{0, 1, 2, ...\}$ may refer to the number of rooms in a house, $x_2^{(i)}$ could be the square feet, etc.
- Unstructured data



Today's Lecture

- Supervised Learning Setup
- Linear Regression
- Logistic Regression ← binary classification
- Computation Graph

Linear Regression

$x^{(i)} \in \mathbb{R}^n \rightarrow y^{(i)} \in \mathbb{R}$

$f_w(x^{(i)}) = w_0 + w_1 x_1^{(i)} + \cdots + w_n x_n^{(i)}$

parameters (weights)

$= w_0 + \sum_{i=1}^{n} w_i x_i \leftarrow \quad \in \mathbb{R}^{n+1}$

$x_0 = 1$ (intercept term) $\quad = \sum_{i=0}^{n} w_i x_i = w^T x$

Linear Regression

$\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}) \cdots (x^{(m)}, y^{(m)})\}$

$\Rightarrow \hat{y} = f(x) \approx y$

$\Rightarrow$ Cost function:

$$J(w) = \frac{1}{2} \sum_{i=1}^{m} (f_w(x^{(i)}) - y^{(i)})^2$$

$f_w(x^{(i)}) = w^T x$

Ordinary least square

Linear Regression

How to choose $\underline{w}$ minimize $J(w)$?

Gradient descent

⇉ initial guess $\underline{w}$.

⇉ $\underline{w} \leftarrow \underline{w} - \beta \nabla_{\underline{w}} J(w)$

β: learning rate / step size.

$w, x \in \mathbb{R}^{n+1}$

$w_j \leftarrow w_j - \beta \frac{\partial}{\partial w_j} J(w)$, $j = 0, \cdots, n$

rate

Linear Regression

Suppose we only have 1 sample

$$J(w) = \frac{1}{2}\left(f(x) - y\right)^2 = \frac{1}{2}\underbrace{\left(w^T x - y\right)^2}_{z}$$

$$\frac{\partial}{\partial w_j} J(w) = \frac{\partial J(w)}{\partial z} \cdot \frac{\partial z}{\partial w_j}$$

$z = \sum_{i=0}^{n} w_i x_i - y$

$w_j x_j$

Chain rule

$$= \left(w^T x - y\right) x_j$$

$f(x)$

$$w_j \leftarrow w_j - \beta\left(f(x) - y\right) x_j$$

Linear Regression

$$w_j \leftarrow w_j - \beta \sum_{i=1}^{m} (f(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

β: learning rate

* Why we use the squared loss in $J(w)$?

$$\Rightarrow y^{(i)} = \theta^T x^{(i)} + \varepsilon^{(i)}$$

$\varepsilon^{(i)}$: error term

Linear Regression

$$\varepsilon^{(i)} \sim N(0, \sigma^2), \quad \forall\, i = 1, \ldots, m$$

independently, identically distributed

$$p(\varepsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\varepsilon^{(i)})^2}{2\sigma^2}\right)$$

$$\Rightarrow \; y^{(i)} \mid \underline{x}^{(i)}; \underline{\theta} \sim N(\theta^T x^{(i)}, \sigma^2)$$

Linear Regression

Given θ, and m data points $x^{(1)} \dots x^{(m)}$

what is the joint distribution of

$y^{(1)} \dots y^{(m)}$?

$$L(\theta) = \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta)$$

$$= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

Linear Regression

Find θ that maximize the likelihood that you see this dataset.

$$\max_{\theta} L(\theta)$$

$$\ell(\theta) = \log L(\theta)$$

$$= \sum_{i=1}^{m} \log \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2} \right) \right]$$

Linear Regression

(continued)

$$= \sum_{i=1}^{m} \left[\log \frac{1}{\sqrt{2\pi}\sigma} - \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2} \right]$$

$$= m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2$$

$$\max_{\theta} L(\theta) \iff \min_{\theta} \sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2$$

Today's Lecture

- Supervised Learning Setup
- Linear Regression
- Logistic Regression
- Computation Graph

Logistic Regression

y take value from 0, 1.

0 → not

1 → spam email

<1> Can we use linear regression model?

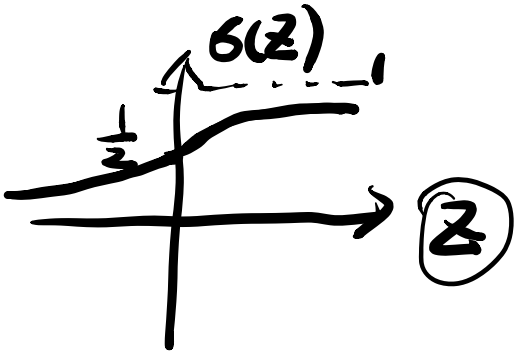
$\hat{y} = w^T x^{(i)}$

- Can be larger than 1 or smaller than 0

Logistic Regression

(2) $\hat{y}^{(i)} = \sigma(w^T x^{(i)})$

$$\sigma(z) = \frac{1}{1+e^{-z}}$$ < sigmoid function
logistic function.



$z \to +\infty \quad \sigma(z) \to 1$

$z \to -\infty \quad \sigma(z) \to 0$

Logistic Regression

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

$$h(x) = \frac{f(x)}{g(x)}$$

$$h'(x) = \frac{g(x)f'(x) - f(x)g(x)}{g(x)^2}$$

$$\sigma'(z) = \frac{0-(-e^{-z})}{(1+e^{-z})^2}$$

$$= \frac{e^{-z}}{(1+e^{-z})^2} = \frac{1}{1+e^{-z}} \cdot \frac{e^{-z}}{1+e^{-z}}$$

$\sigma(z)$ $\quad [1-\sigma(z)]$

Logistic Regression

$$\hat{y} = \sigma(w^T x) = f(x) \quad [0, 1]$$

$$\begin{cases} P(y=1 \mid x) = \sigma(w^T x) = f(x) \\ P(y=0 \mid x) = 1 - \sigma(w^T x) = 1 - f(x) \end{cases}$$

$$P(y \mid x) = [f(x)]^y [1 - f(x)]^{1-y}$$

Logistic Regression

$$L(w) = \prod_{i=1}^{m} P(y^{(i)} \mid x^{(i)})$$

$$= \prod_{i=1}^{m} f(x^{(i)})^{y^{(i)}} [1 - f(x)]^{1-y^{(i)}}$$

$$l(w) = \log L(w) = \sum_{i=1}^{m} \log \frac{f(x^{(i)})^{y^{(i)}}}{[1 - f(x^{(i)})]^{1-y^{(i)}}}$$

Logistic Regression

$$= \sum_{i=1}^{m} \left[y^{(i)} \log f(x^{(i)}) + (1 - y^{(i)}) \log (1 - f(x^{(i)})) \right]$$

cross-entropy loss

$$w \leftarrow w + \beta \nabla_w L(w)$$

β: learning rate; $\nabla_w L(w)$: gradient

Logistic Regression

$$w_j \leftarrow w_j + \beta \cdot \frac{\partial l(w)}{\partial w_j}$$

Chain rule

$$\frac{\partial l(w)}{\partial w_j} = \frac{\partial l(w)}{\partial f} \cdot \frac{\partial f}{\partial z} \cdot \frac{\partial z}{\partial w_j}$$

Logistic Regression

(continued)

$$= \sum_{i=1}^{m} \left(y^{(i)} \frac{1}{f(x^{(i)})} - (1 - y^{(i)}) \frac{1}{1 - f(x^{(i)})} \right]$$

$$\sigma(z) \cdot (1 - \sigma(z)) \, x_j^{(i)}$$

$$= \sum_{i=1}^{m} \Big(y^{(i)} [1 - f(x^{(i)})] - (1 - y^{(i)}) f(x^{(i)}) \Big] x_j^{(i)}$$

Logistic Regression

$$= \sum_{i=1}^{m} \left[y^{(i)} - f(x^{(i)}) \right] x_j^{(i)}$$

linear regression,

$$f(x^{(i)}) = w^T x^{(i)}$$

Logistic regression

$$f(x^{(i)}) = \sigma(w^T x^{(i)})$$

Today's Lecture

- Supervised Learning Setup
- Linear Regression
- Logistic Regression
- Computation Graph

Computation Graph

$J(a,b,c) = 2(a+bc)$

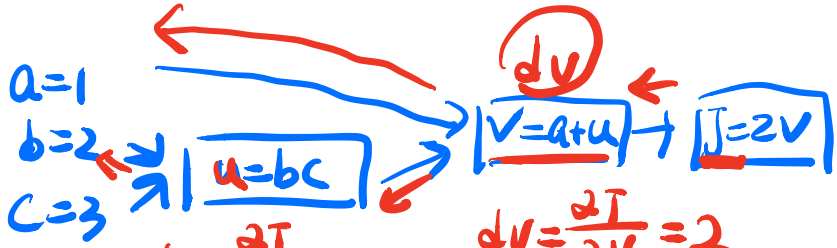
$= 2\times(1+6) = 14$

$a=1$
$b=2$
$c=3$

$u = bc$

$v = a+bc = a+u$

$J = 2v$



- $du = \frac{\partial J}{\partial u} = \frac{\partial J}{\partial v}\cdot\frac{\partial v}{\partial u} = 2\cdot 1 = 2$
- $db = \frac{\partial J}{\partial b} = \frac{\partial J}{\partial u}\cdot\frac{\partial u}{\partial b} = 2\cdot c = 2\cdot 3 = 6$