CS 229 Lecture 1 - Introduction

Lecture 2 - Linear Regression and Gradient Descent

- Hypothesis $h(x) = \sum_{j=0}^{n} \theta_j x_j$

  where $x_0 = 1$ $\quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} \quad x = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix}$

  θ = "parameters"

  m = # training examples

  x = features

  y = target $\Rightarrow$ (x, y) : training example

  $(x^{(i)}, y^{(i)})$ = i^{th} training example

  n = # features

→ How to choose θ

$h(x) \approx y \approx h_\theta(x)$

minimize $\frac{1}{2} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{i} \right)^2 =: J(\theta)$

- Gradient Descent

  Start with $\theta = \vec{0}$

  $\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$ (α: learning rate)

  $a := a + 1$

→ $\frac{\partial}{\partial \theta_j} J(\theta) = \left(h_\theta(x) - y \right) \cdot \frac{\partial}{\partial \theta_j} \left(h_\theta(x) - y \right)$

$$= (h_\theta(x) - y) \cdot x_j$$

$$\Rightarrow \theta_j := \theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)}$$

Repeat until convergence

(for $j = 1 \dots n$)

→ Batch gradient descent → Compute whole Training Sample
↘ Slower but accurate

→ Stochastic gradient → much faster
↘ compute a single Training Sample

• Normal equation $\theta \in \mathbb{R}^{n+1}$

$$\nabla_\theta J(\theta) = \begin{bmatrix} \frac{\partial J}{\partial \theta_0} \\ \dots \\ \frac{\partial J}{\partial \theta_n} \end{bmatrix}$$

↳ $\nabla_\theta J(\theta) \overset{\text{set}}{=} \vec{0}$

$A \in \mathbb{R}^{m \times n}$

$$\nabla_A f(A) = \begin{bmatrix} \frac{\partial f}{\partial A_{11}} & \dots & \frac{\partial f}{\partial A_{1n}} \\ \vdots & & \\ \frac{\partial f}{\partial A_{m1}} & \dots & \frac{\partial f}{\partial A_{mn}} \end{bmatrix}$$

$\text{tr} A$ = trace of A $\sum_i^n A_{ii}$

$\Rightarrow \text{tr} A = \text{tr} A^T$

$\nabla \rightarrow A_{ij} = A_{ji}$ but $i = j$

• exam: $A \in \mathbb{R}^{2 \times 2}$

$f(A) = A_{11} + A_{12}^2$

$\Rightarrow \nabla_A f(A) = \begin{bmatrix} 1 & 2A_{12} \\ 0 & 0 \end{bmatrix}$

$f(A) = \text{tr} AB$ ↳ fixed Matrix

↳ $\nabla_A f(A) = B^T$

$\text{tr} AB = \text{tr} BA$

$\text{tr} ABC = \text{tr} CAB$

$\nabla_A \text{tr} A A^T C = CA + C^T A$

↳ $\frac{d}{da} a^2 c = 2ac$

Review these!

Lecture linear + Notes

- elementwise operations $X\theta$

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}\left(h(x^{(i)}) - y^{(i)}\right)^2$$

$$= \frac{1}{2}(X\theta - y)^T(X\theta - y) \quad \text{Error terms}$$

$$z^T z = \sum_i z_i^2$$

$$\nabla_\theta J(\theta) = \nabla_\theta \frac{1}{2}(X\theta - y)^T(X\theta - y)$$

$$= X^T X\theta - X^T y \overset{\text{set}}{=} 0$$

$$\Rightarrow \theta = (X^T X)^{-1} X^T y \quad \text{(Normal function)}$$

↳ if X is non-invertible → redundant feature

Lecture 3: locally weighted & logistic Regression

Recap:

$(x^{(i)}, y^{(i)})$ - i^{th} example

$x^{(i)} \in \mathbb{R}^{n+1}$, $y^{(i)} \in \mathbb{R}$, $x_0 = 1$

m = # examples, n = # features

$$h_\theta(x) = \sum_{j=0}^{n} \theta_j x_j = \theta^T x$$

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

→ locally weighted Regression

"parametric" learning algorithm

↳ fit fixed set of θ_i to data

"Non-parametric" learning algorithm

↳ Amount of data / θs you need to keep grows (linearly) with size of data

→ LR: fit θ to minimize

$$\frac{1}{2}\sum_i^m \left(y^{(i)} - \theta^T x^{(i)}\right)^2$$

Return $\theta^T x$

→ locally weighted Regression: fit θ to minimize

$$\sum_{i=1}^m w^{(i)} \left(y^{(i)} - \theta^T x^{(i)}\right)^2$$

where $w^{(i)}$ is a weighting func

$$w^{(i)} = \exp\left(-\frac{(x^{(i)} - x)^2}{2\tau^2}\right)$$

if $|x^{(i)} - x|$ is small, $w^{(i)} \approx 1$

$(x^{(i)}, y^{(i)})$ if $|x^{(i)} - x|$ is large, $w^{(i)} \approx 0$

↳ how much you should pay attention to $x^{(i)}$

τ = "bandwidth"

→ Abbreviation
i.e. id est
e.g. for example

- Probabilistic interpretation

Why least squares?

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

↑ "error": unmodeled effects, random noise

$$\varepsilon^{(i)} \propto \mathcal{N}(0, \sigma^2)$$

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

→ this implies ("parameterized by" θ, not random; mean ↑)

$$p(y^{(i)} \mid x^{(i)}; \theta) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

I.e. $y^{(i)} \mid x^{(i)}, \theta \propto \mathcal{N}(\theta^T x^{(i)}, \sigma^2)$ → variance

$$L(\theta) = p(\vec{y} \mid X; \theta)$$ → prob: θ fixed, vary data

"likelihood of θ" $= \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta)$

↓ training fixed, vary θ

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

- likelihood of parameters theta
- probability of data

"log likelihood"

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

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

Gaussian IID

MLE: maximum likelihood estimate

Choose θ to maximize $L(\theta)$

i.e. choose θ to minimize: $\frac{1}{2} \sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2 = J(\theta)$

Classification problem

$y \in \{0, 1\}$ (binary classification)

→ Logistic Regression

Want $h_\theta(x) \in [0,1]$ "sigmoid" or "logistic" function

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

$$\begin{cases} p(y=1 \mid x; \theta) = h_\theta(x) \\ p(y=0 \mid x; \theta) = 1 - h_\theta(x) \end{cases}$$

$$\hookrightarrow p(y \mid x; \theta) = h(x)^y (1 - h(x))^{1-y}$$

$$L(\theta) = p(\vec{y} \mid x; \theta)$$

$$= \prod_{i=1}^{m} h(x^{(i)})^{y^{(i)}} (1 - h(x^{(i)}))^{1-y^{(i)}}$$

$$l(\theta) = \sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))$$

→ Choose θ to $\max l(\theta)$

Batch gradient descent

$$\theta_j := \theta_j + \alpha \frac{\partial}{\partial \theta_j} l(\theta)$$

$$= \theta_j + \alpha \sum_{i=1}^{m} \left[y^{(i)} - h_\theta(x^{(i)}) \right] \cdot x_j^{(i)}$$

$$= -\alpha \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) \cdot x_j^{(i)}$$

Newton's Method

Have f, want to find θ

s.t: $f(\theta) = 0$ (i.e want $l'(\theta) = 0$

$$\theta^{(t+1)} := \theta^{(t)} - \frac{f(\theta^{(t)})}{f'(\theta^{(t)})}$$ "Quadratic convergence"

(let $f(\theta) = l'(\theta)$)

$$\rightarrow \theta^{(t+1)} := \theta^{(t)} - \frac{l'(\theta^{(t)})}{l''(\theta^{(t)})}$$

- when θ is a vector: $\theta \in \mathbb{R}^{n+1}$ $\xrightarrow{}$ $\mathbb{R}$ $\quad n+1 \times n+1$

$$\theta^{t+1} := \theta^{(t)} + H^{-1} \underbrace{\nabla_\theta l}_{\rightarrow \text{vector } \mathbb{R}^{n+1}}$$

where H is the Hessian matrix

$$H_{ij} = \frac{\partial^2 l}{\partial \theta_i \partial \theta_j}$$ $\rightarrow$ expensive

Lecture 4 Perceptron & Generalized Linear Model

Logistic Regression Sigmoid

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

$$g(z) = \begin{cases} 1, & z \geq 0 \\ 0, & z < 0 \end{cases}$$

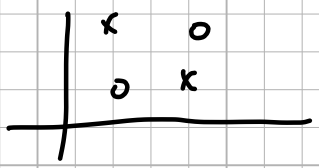
$$h_\theta(x) = g(\theta^T x)$$

$$= \frac{1}{1+e^{-\theta^T x}}$$

$$\theta_j := \theta_j + \alpha \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}$$

(not used in practice)

e.g.



- Exponential Family

PDF: $p(y;\eta) = b(y) \exp\left[\eta^T T(y) - a(\eta)\right]$

y - data

η - eta: natural parameter

$T(y) \overset{=y}{!}$ - sufficient statistic ?

$b(y)$ - base measure

$a(\eta)$ - log partition

- Bernoulli (Binary Data)

ϕ : prob of event

$p(y;\phi) = \phi^y (1-\phi)^{(1-y)} = \exp\left(\log\left(\phi^y (1-\phi)^{1-y}\right)\right)$

$= \exp\left(\log\left(\frac{\phi}{1-\phi}\right) \cdot y + \log(1-\phi)\right)$

$\frac{\phi^y \cdot (1-\phi)}{(1-\phi)^y}$

$$\begin{cases} b(y) = 1 \\ \eta = \log\left(\frac{\phi}{1-\phi}\right) \Rightarrow \phi = \frac{1}{1+e^{-\eta}} \\ T(y) = y \\ a(\eta) = -\log(1-\phi) \Rightarrow -\log\left(1 - \frac{1}{1+e^{-w}}\right) = \log(1+e^w) \end{cases}$$

- Gaussian (with fixed variance)

Assume $\sigma^2 = 1$

$$P(y;\mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2}\right)$$

$$= \underbrace{\frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}}_{b(y)} \exp\left(\underbrace{\mu}_{\eta} y - \underbrace{\frac{1}{2}\mu^2}_{a(\eta)}\right)$$

Properties:

@ MLE wrt η $\Rightarrow$ Concave

NLL is convex

① $E[y;\eta] = \frac{\partial}{\partial \eta} a(\eta)$

② $\text{Var}[y;\eta] = \frac{\partial^2}{\partial \eta^2} a(\eta)$

} No integrals

- GLM

Assumption / Design choices

(i) $y \mid x;\theta \sim$ Exponential Family (η)

(ii) $\eta = \theta^T x \quad \theta \in \mathbb{R}^w, x \in \mathbb{R}^w$

(iii) Test time: output $E[y \mid x;\theta] = h_\theta(x)$

$= E[y;\theta^T x]$

mean is the predict

↳ $\max_\theta \log p(y^{(i)}; \theta^T x^{(i)})$

GLM training

→ learning update rule

$$\theta_j := \theta_j + \alpha (y^{(i)} - h_\theta(x^{(i)}))\, x_j^{(i)}$$

Terminology

η : natural parameter

$E[y; \eta] = g(\eta)$ → Canonical Response function

$\eta = g^{-1}(u)$ → Canonical link function

$$g(\eta) = \frac{\partial}{\partial \eta} a(\eta)$$

- $\exists$ - parameterizations

Model Param	Natural Param	Canonical Para
θ	η	ϕ - Ber
		μ, σ^2 : Gauss
↑	$\xrightarrow{g}$	λ : Poisson
learn - $\theta^T x$ →	$\xleftarrow{g^{-1}}$	
Design choice		

Real - Gaussian

Binary - Bernoulli

Count - Poisson

$\mathbb{R}^+$ - Gamma, Exponential

Prctn - Beta, Dirichlet

Bayesian

- Logistic Regression : $h_\theta(x) = E[y | x; \theta] = \phi = \frac{1}{1+e^{-\eta}} = \frac{1}{1+e^{-\theta^T x}}$

• Gaussian $\eta = \mu$

• Soft max Regression - Cross Entrophy

→ Multiclassification

k: # classes label $y \in \{0,1\}^k$ one-hot vector

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

$\hat{p}(y)$ - $p(y)$

$$\text{Cross Ent}(p, \hat{p}) = -\sum_{y \in \{0,1,\dots\}} p(y) \cdot \log \hat{p}(y)$$

$$= -\log \hat{p}(y_\Delta)$$

$$= -\log \frac{e^{\theta_\Delta^T x}}{\sum_{c \in \{\Delta,\square,\circ\}} e^{\theta_c^T x}}$$

Lecture 5 GDA & Naive Bayes

• Discriminative Learning Algorithms

• Gaussian Discriminant Analysis

• Generative & Discriminative comparison

• Naive Bayes

$x \to y$

→ Discriminative: learn $p(y|x)$

(or learn $h_\theta(x) = \begin{cases} 0 \\ 1 \end{cases}$ directly)

→ Generative learning algo:

Learns $P(x|y)$ (x: feature, y: class) $\quad P(y)$ → class prior

- Bayes rule:

$$P(y=1|x) = \frac{P(x|y=1) \cdot p(y=1)}{P(x)}$$

$$p(x) = p(x|y=1)\, p(y=1) + P(x|y=0) \cdot p(y=0)$$

- Gaussian Discriminal Analysis (GDA)

Suppose: $x \in \mathbb{R}^n$ (drop $x_0 = 1$ convention)

Assume $p(x|y)$ is Gaussian

d: dimension of matrix

[multivariate Gaussian $f(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$

uni-variate]

$z \sim \mathcal{N}(\vec{\mu}, \Sigma) \quad z \in \mathbb{R}^n$

$\vec{\mu} \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{n \times n}$

→ $E[z] = \mu$

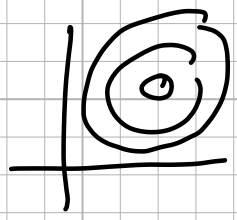
→ $\mathrm{Cov}(z) = E[(z-\mu)(z-\mu)^T]$

$= E\, zz^T - (Ez)(Ez)^T$

→ $p(z) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(\frac{-1}{2} \cdot (x-\mu)^T \Sigma^{-1} (x-\mu)\right)$

e.g. $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ Contour of Gaussian density

$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$



- GDA model. Parameters: $\mu_0, \mu_1, \Sigma, \phi$

$p(x|y=0), p(x|y=1)$ $\quad \mathbb{R}^n \quad \mathbb{R}^{n \times n} \quad \mathbb{R}$

$p(y) = \phi^y (1-\phi)^{1-y}$

→ training set $\{(x^{(i)}, y^{(i)})\}_{i=1}^m$

Joint likelihood:

$$\mathcal{L}(\phi, \mu, \mu_1, \Sigma) = \prod_{i=1}^m p(x^{(i)}|y^{(i)}) \cdot p(y^{(i)})$$

Discrimination:

Conditional likelihood

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

Maximize likelihood estimation

$$\phi = \frac{\sum_{i=1}^m y^{(i)}}{m} = \frac{\sum_{i=1}^m 1\{y^{(i)}=1\}}{m}$$

$$\mu_0 = \frac{\sum_{i=1}^m 1\{y^{(i)}=0\} x^{(i)}}{\sum_{i=1}^m 1\{y^{(i)}=0\}}$$

Tương tự μ_1

$$\Sigma = \frac{1}{m} \sum_{i=1}^{m} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T$$

→ Prediction [$\min_z (z-5)^2 = 0$ $\arg\min_z (z-5)^2 = 5$]

$$\arg\max_y p(y|x) = \arg\max_y \frac{p(x|y)\,p(y)}{p(x)}$$

$$= \arg\max_y p(x|y)\,p(y)$$

Generative ⇒ Discriminative

Gaussian — Logistic Regression

(stronger) — (weaker)

→ Cyclical Test and gain conviction

or just use Logistic Regression

→ Naive Bayes

feature vector x? $x = \begin{bmatrix} \\ \\ \\ \end{bmatrix}$ $x \in \{0,1\}^n$

$x_i = 1$ {word i appears in mail}

2^{10000} possible value

- Assum $x_i \overset{n}{\Sigma}$ are conditional independent

$$p(x|y) = \prod_{i=1}^{n} p(x_i|y)$$

- Parameters $\phi_{j|y=1} = p(x_j = 1 | y = 1)$

$\phi_{j|y=0} = p(x_j = 1 | y = 0)$

$\phi_y = p(y = 1)$

Joint likelihood:

$$\mathcal{L}(\phi_y, \phi_{j|y}) = \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}; \phi_y, \phi_{j|y})$$

MLE:

$$\phi_y = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}{m}$$

$$\phi_{j|y=1} = \frac{\sum_{i=1}^{m} 1\{x_j^{(i)} = 1, y^{(i)} = 1\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}$$