কৃত্রিম বুদ্ধিমত্তা - Artificial intelligence মেশিন লার্নিং - Machine learning

মেশিন লার্নিং (৩): কস্ট ফাংশনের অন্তরীকরণ এবং গ্র্যাডিয়েন্ট ডিসেন্ট

Sharif Hasan • November 6, 2020 সর্বশেষ আপডেট January 8, 2021 3 819 পড়তে 5 মিনিট লাগতে পারে

Andrew Ng এর মেশিন লার্নিং কোর্সে গ্র্যাডিয়েন্ট ডিসেন্ট সম্পর্কে চমৎকার ব্যাখ্যা করা আছে। আমার এই লেখায় অনেক কিছুই তার লেকচার থেকে অনুপ্রাণিত হয়ে লিখা। আজকের লিখায় কস্ট ফাংশনের অন্তরীকরণ এবং গ্র্যাডিয়েন্ট ডিসেন্ট (Gradient descent) নিয়ে আলোচনা করার চেষ্টা চালাবো।

যারা পূর্বের লিখা গুলো পড়েননি তারা আমার আগের দুটি লিখা পরে নিতে পারেন। আগের লিখার আমরা (১) মেশিন লার্নিং এর শুরু এবং, (২) লিনিয়ার রিগ্রেশন নিয়ে কথা বলেছিলাম। আজকের লিখায় আমরা আমাদের কস্ট ফাংশনের অন্তরীকরণ করবো এবং এই অন্তরক আমরা গ্র্যাডিয়েন্ট ডিসেন্ট এ ব্যাবহার করে আমাদের কস্ট ফাংশনের মান কিভাবে কমানো যায় তা নিয়ে হাল্কার উপর ঝাপসা আলোচনা করবো।

এই লেখাটি শুরুর আগে অমি ধরে নিলাম আপনাদের আগে থেকেই লিনিয়ার রিগ্রেশন নিয়ে প্রাথমিক ধারনা আছে :)।

ফিরে দেখা...

লিনিয়ার রিগ্রেশন এর জন্য আমাদের একটি লিনিয়ার হাইপথিসিস ফাংশন $h(x) = \theta_0 + \theta_1 x$ আছে। এখন আমাদের কাজ হলো θ_0 এবং θ_1 এর জন্য আমাদের এমন মান বের করতে হবে যার জন্য $h(x)$ রেখাটি আমাদের ট্রেইনিং সেট এ বেস্ট ফিট করে। আমাদের ট্রেইনিং সেটে দুটি কলাম আছে এবং যারা x,y দিয়ে লেবেল করা। এখানে x হলো ইনপুট এবং y হলো আউটপুট। টেবিলের i তম রো (row)/ ট্রেইনিং উদাহরণ $x^{(i)}, y^{(i)}$ দিয়ে লেবেল করা। এটা কিন্তু কোনও সূচক না। এটা শুধু বুঝায় যে এটা i তম রো এর ট্রেইনিং ডেটা/ উদাহরণ।

MSE কস্ট ফাংশন

প্যারামিটার θ এর কোন মানের জন্য আমাদের MSE (**M**ean **S**quared **E**rror) কস্ট ফাংশন J হবে,

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

$$\overline{i=1}$$

এখানে যে টার্ম গুলো ব্যাবহার করেছি তাদের মানে নিচে দেখি ,

m	টেবিল এ মোট ডাটা এর সংখ্যা (টেবিল এর মোট রো এর পরিমাণ)
$x^{(i)}$	i তম রো এর ইনপুট ভেক্টর (X লেবেল করা কলাম)
$y^{(i)}$	i তম রো এর আউটপুট ভেক্টর (y লেবেল করা কলাম)
θ	প্যারামিটার গুলোর মান, যার জন্য আমাদের হাইপথিসিস ফাংশন ডাটাসেট এ বেস্ট ফিট হয় ($\theta_0, \theta_1, \theta_2$ etc.)
$h_\theta(x^{(i)})$	i তম ট্রেইনিং ইনপুট ডেটা এবং প্যারামিটার θ_k ব্যাবহার করে আমাদের এলগরিদম এর প্রেডিকশন

আমাদের **MSE** কস্ট ফাংশন তাহলে কি পরিমাপ করছে**?**

MSE কস্ট ফাংশন আমাদের মডেলের আউটপুট গড়ে কতটুকু টার্গেট আউটপুট থেকে দুরে আছে তা পরিমাপ করে। সুতরাং আমরা ভাবতেই পারি এই কস্ট ফাংশন আসলে ট্রেইনিংসেটে আমাদের এলগরিদম/মডেল এর পারফরমান্স হিসাব করে। যদি কস্ট ফাংশনের মান যদি বেশি হয় তাহলে বুঝা যায় আমাদের এলগরিদম ট্রেইনিংসেটে ভাল ভাবে পারফর্ম করতে পারছে না (বেস্ট ফিট হয়নি)। এখন আমাদের লার্নিং এলগরিদম এর উদ্দেশ্য হলো প্যারামিটার $\boldsymbol{\theta}$ এর মান কে এমনভাবে সেট করা যাতে আমাদের কস্ট ফাংশনের মান সর্বনিম্ন হয়।

এটার মানে একসময় বুঝা যাবে 🙂

এক চলকের গ্র্যাডিয়েন্ট ডিসেন্ট, কস্ট মিনিমাইজ করা

আমরা গ্র্যাডিয়েন্ট ডিসেন্ট ব্যাবহার করে θ এর এমন মান খুঁজে বের করবো যাতে আমাদের MSE কস্ট ফাংশন J এর মান সর্বনিম্ন করা সম্ভব হয়। বুঝার জন্য আমরা আপাতত MSE কে ভুলে যাই এবং $J(\theta) = \theta^2$ ধরে নিই।

এখন বলার চেষ্টা করি θ এর কোন মান এর জন্য আমাদের $J(\theta)$ এর মান সর্বনিম্ন হবে।

গ্র্যাডিয়েন্ট ডিসেন্ট সাধারণত থেটার যেকোনো রেন্ডম মান নিয়ে স্টার্ট হয়। তারপর আস্তে আস্তে মিনিমাম কস্ট ফাংশনের মিনিমাম এর দিকে ধাবিত হয়। এক্ষেত্রে $\theta = 0$ হলে আমাদের $J(\theta)$ এর মান সর্বনিম্ন হবে। ধরুন আমরা $\theta = 3$ থেকে শুরু করেছি।

গ্র্যাডিয়েন্ট ডিসেন্ট একটা ইটারেটিভ এলগরিদম। প্রত্যেকটা ইটারেশন এ আমরা নিম্নরূপে আমাদের প্যারামিটার θ আপডেট করবো,

> *আপনি যদি এর আগে $:=$ চিহ্ন দেখে না থাকেন, এই চিহ্ন দ্বারা বুঝায় সমীকরণ এর ডান পাশের মান বামপাশের চলকে এসাইন করা হয়েছে।*

এখানে α কে বলা হয় লার্নিং রেট। এর মান আমাদের নিজ হাতে নির্বাচন করে দিতে হয়। একে বলা হয় হাইপারপারামিটার। আমরা এখানে পরে আসবো। এখনকার জন্য আমরা একে **0.1** সেট করে দিচ্ছি। θ এর সাপেক্ষে $J(\theta)$ এর অন্তরক হল 2θ.

নিচে নামার আগে আমরা θ বনাম $J(\theta)$ এর প্লটটি একটু খেয়াল করে নিবো। লক্ষ করলে দেখছি আমাদের কার্ভটি একটি অবতল বক্ররেখা যার $\theta = 0$ বিন্দুতে সর্বনিম্ন মান অবস্থিত।

এখন আমরা দেখি θ এর কোন মানের জন্য আমাদের কস্ট ফাংশনের মান সর্বনিম্ন হয়, নিচের টেবিলটি লক্ষ করি।

উপরের টেবিল এ আমাদের ১০ টি ইটেরাশন এর জন্য থেটা এর মান এর পরিবর্তন দেখানো হয়েছে। আমরা প্রথম ইটেরেশন এর $\theta = 3$ থেকে শুরু করেছি। তখন $\alpha \frac{d}{d\theta} J(\theta)$ এর মান ছিলো 0.6। সুতরাং এই ধাপে θ এর পরিবর্তিত মান হবে

$\theta := 3 - 0.6 = 2.4$। এভাবে আমরা পরবর্তী স্টেপ গুলো দেখলে দেখতে পাই আমাদের θ র মান আস্তে আস্তে ০ এর দিকে অগ্রসর হচ্ছে। অর্থাৎ আমাদের কস্ট ফাংশনের মান কমছে।

কস্ট ফাংশনের ডেরিভেটিভ (অন্তরক)

গ্র্যাডিয়েন্ট ডিসেন্ট কেন কস্ট ফাংশনের অন্তরক ব্যাবহার করে? অন্তরকের মাদ্ধমে আমরা বর্তমান θ এর পজিশন এ কস্ট ফাংশনের ঢাল বের করি।

এই অন্তরকের মান আমাদের দুইটা কথা বলে দেয়।

প্রথম কথা এই অন্তরক আমাদের বলে দিতে পারে আমরা কোন দিকে আমাদের θ কে সরিয়ে নিবো। মানে আমাদের বলে দেয় আমাদের θ এর মান কমবে না বাড়বে।

যদি কস্ট ফাংশনের মান ধনাত্মক হয় তাহলে আমরা বলতে পারি কস্ট ফাংশনের বাম দিকে সর্বনিম্ন বিন্দু অবস্থান করছে। সুতরাং আমাদের θ এর মান কমাতে হবে যদি আমরা কস্টফাংশন কে মিনিমাম করতে চাই।

আবার যদি আমরা ঋণাত্মক কস্ট ফাংশন পাই, তাহলে বলতে পারি আমরা ডান দিকে গেলে কস্ট ফাংশনের মিনিমাম পাব। শতরাং আমাদের θ এর মান বারাতে হবে যদি আমরা কস্ট ফাংশনের সর্বনিম্ন খুঁজে পেতে চাই।

দ্বিতীয় কথা, এই অন্তরক এর মান আমাদের বলতে পারে θ এর মান আমরা কি হারে বাড়াবো বা কমাবো (Step size)। আমরা যখন মিনিমাম থেকে অনেক দুরে থাকবো তখন আমাদের বড়ো বড়ো স্টেপ নিতে হবে।কারণ তখন ঢাল এর মান অনেক বেশি হবে। বড়ো স্টেপ এর ফলে আমরা দ্রুত মিনিমামের দিকে আগাতে পারি। যখন আমাদের

ক্ষুদ্র স্টেপ নিতে হবে তখন আমাদের ঢাল এর মান কম হবে। এভাবে আমরা যত কাছে যাবো আমদের স্টেপ সাইজ তত কমতে থাকবে। অর্থাৎ প্রতি ইটেরেশন আমাদের কস্ট ফাংশন মিনিমামের দিকে যেতে থাকবে এবং মিনিমামের যত কাছে যাবে আমাদের স্টেপ সাইজ তত কম হতে থাকবে।

লার্নিং রেট (আলফা) [The Learning Rate – Alpha]

লার্নিং রেট মেশিন লার্নিং ইঞ্জিনিয়ারের উপর কিছু বাড়তি কন্ট্রোল দেয় যাতে আমাদের স্টেপ সাইজ কত বড়ো বা ছোট হবে তা নিয়ন্ত্রণ করেতে পারি। সঠিক লার্নিং রেট নির্ধারণ করা একটু জটিল বিষয়। আপনি যদি আলফার মান খুব বেশি সেট করেন তাহলে আপনার এলগরিদম দরকার এর চেয়ে বেশি বড় স্টেপ নিতে পারে। আবার আপনি যদি খুব কম নেন তাহলে আপ্নাল মডেল লার্ন করতে অনেক সময় নিবে।

উদাহরণসরূপ আমরা উপরের উদাহরনে লার্নিং রেট ২ সেট করে দেখতে পারি। তাহলে দেখবো প্রতিটি ইটেরেশন আমাদের মিনিমাম থেকে দুরে নিয়ে যাবে।

গ্র্যাডিয়েন্ট ডিসেন্ট বন্ধ করা

উপরের উদাহরণে দেখার বিষয় হলো আসলে কখন গ্র্যাডিয়েন্ট ডিসেন্ট আমাদের কস্ট ফাংশনকে শুন্য করবে না, অর্থাৎ $\theta = 0$ হবে না। সুতরাং আমরা যদি আমাদের টার্গেট আউটপুটের খুব কাছাকাছি যেতে পারি তবে আমরা আমাদের গ্র্যাডিয়েন্ট ডিসেন্ট বন্ধ করে দিতে পারবো।

একাধিক চলক এর গ্র্যাডিয়েন্ট ডিসেন্ট।

MSE কস্ট ফাংশনে আমরা একাধিক চলক ব্যাবহার করতে পারি। তার আগে আমরা উপরের মত আরেকটি উদাহরণ দেখবো, তবে এবার একাধিক চলক ব্যাবহার করে। সুতরাং ধরে নিই,

ফাংশন, $J(\theta) = \theta_1^2 + \theta_2^2$

যখন আমাদের একাধিক চলক থাকবে, প্রত্যেকটি চলকের জন্য আমাদের আপডেট রুল আলাদা আলাদা হবে। θ_1 এর জন্য আমাদের আপডেট রুল θ_1 এর সাপেক্ষে $J(\theta)$ এর আংশিক অন্তরক ব্যাবহার করবে। একইভাবে আমরা θ_2 এর জন্য θ_2 এর সাপেক্ষে $J(\theta)$ এর আংশিক অন্তরক ব্যাবহার করবো। নিচের চিত্রে আমরা আপডেট রুল এর সাথে সম্পর্কিত ক্যালকুলাস ও দেখতে পারছি।

আশা করি বুঝা যাচ্ছে আমরা কি করতে চাচ্ছি। আমাদের আপডেট রুল এ আমরা একচলকের মত করেই আপডেট করেছি। তবে এখানে আমাদের দুটি চলক আছে। তাই দুটিকেই আলাদা করে আপডেট করতে হচ্ছে। এখানে একটা বিষয় নোট করা উচিত যে, যদি কখন কোনোভাবে (অনেক হবে আসলে) $\boldsymbol{\theta_1}$ এর উপর $\boldsymbol{\theta_2}$ এর মান নির্ভর করে তবে আমরা একই স্টেপে $\boldsymbol{\theta_1}$ এর আপডেট করা মান ব্যাবহার করবো না। আলাদা স্টেপ হলে করা যাবে। আশা করি বুঝা গেছে। না বুঝলে সমস্যা নাই। ইমপ্লেমেন্টেশনের সময় অনেক কিছু পরিষ্কার হয়ে যাবে।

MSE কস্ট ফাংশনের গ্র্যাডিয়েন্ট ডিসেন্ট

এতক্ষণ ধরে আমরা দেখলাম কিভাবে আমরা একটি ফাংশনে গ্র্যাডিয়েন্ট ডিসেন্ট প্রয়োগ করে তার সর্বনিম্ন মান পেতে পারি। এখন সময় হল আমাদের আলোচিত কস্ট ফাংশনে গ্র্যাডিয়েন্ট ডিসেন্ট প্রয়োগ করা।

নিচের [1.0] ইকুয়েশন টি আমাদের MSE কস্ট ফাংশন।

এই সমীকরণের অন্তরক বের করা একটু জটিল। এক্ষেত্রে আমাদের জেনে রাখতে হবে যে $\boldsymbol{x, y}$ কিন্তু আমাদের অন্তরীকরণে চলক নয়, বরং তারা একটি ধ্রুবকের সেট। তাই অন্তরীকরণের সময় আমরা $\boldsymbol{x, y}$ অন্য সাধারণ ধ্রুবকের মতই দেখবো।
পুনশ্চঃ লিনিয়ার রিগ্রেশনের জন্য আমাদের হাইপোথিসিস ফাংশন
$$\boldsymbol{h(x) = \theta_0 + \theta_1 x}$$

নিচে আমাদের কস্ট ফাংশনের অন্তরীকরণ দেয়া হলো।

উপরের হিসাবটি লক্ষ করি। কাজ করার সময় আমরা m কে ২ দ্বারা গুন করেছি। এটি একটি সিম্পল পরিবর্তন। একে বলা হয় **One Half Mean Squared Error**। যার জন্য শেষ এ আমাদের সূচক এর জন্য যে ২ সামনে আসে, ওই দুই ক্যানসেল হয়ে যায়।

কিছু কথা

এখানে লক্ষণীয় যে, আমাদের θ তে প্রতিটি আপডেট, আমাদের ডাটাসেটের গড়ের উপর নির্ভর করে। ট্রেইনিং সেটের প্রতিটি উদাহরণ আমাদের θ এর উপর আলাদা আলদা আপডেট নির্ধারণ করে। তাই আমরা প্রতিটি আপডেট এর গড় নিয়ে শেষ আপডেট করবো। একে বলা হয় **Batch Gradient Descent** এবং প্রতিটি রো যেই আপডেট নির্ধারণ করে তাকে বলা হয় **stochastic gradient descent**।

আজকের লেখাটির অনেকটুকু অনুবাদ করা হয়েছে। আরও কিছু বিষয় + ছবি ইন্টারনেট থেকে সংগৃহীত। লেখায় কোনও ভুল পেলে ক্ষমাসুন্দর চোখে দেখার অনুরোধ করছি। কমেন্ট বা ইমেইল এ মতামত দিতে ভুলবেন না।

লেখাটি কেমন লেগেছে আপনার?

রেটিং দিতে হার্টের উপর ক্লিক করুন।

গড় রেটিং 5 / 5. মোট ভোট: 5

#মেশিন লার্নিং

3 টি মন্তব্য

guEsS

November 9, 2020 at 2:10 PM

Will eagerly wait for the next part..

Reply

Sharif Hasan

November 9, 2020 at 4:32 PM

I will try my best (Mr/Ms/Mrs) 💕

Reply

Faysal Hasan

May 24, 2021 at 9:20 AM

We need next part

Reply