# [ML (Một lằn) (notion.site)](https://quill-agreement-84d.notion.site/ML-M-t-l-n-4b076ae6e8754d4898f170707a763b24)

[**Chapter 1 5**](#_1s86cwnrptxe)

[Học máy là gì? 5](#_57befngi2d1a)

[Học từ data (learnt from data) là gì? 5](#_uc2fyt7kx2ib)

[Model là gì? 5](#_ro0vzr7ijvsd)

[Learning algorithm là gì? 5](#_iqfstmjqig5x)

[3 phần quan trọng nhất của học máy là gì? 5](#_5zzlry90md1r)

[**Chapter 2 5**](#_y2q63fvxlg52)

[Supervised, Unsupervised, Semi-supervised 5](#_h16xvk80fhl9)

[Training data được tạo ra như thế nào? 5](#_221qxdvn3ixx)

[Regression vs Classification 5](#_9t2ttc1rs2c)

[Numerical and categorical variables 5](#_fx6jmwpvmu2n)

[Ví dụ dạng numerical data 5](#_18392sixp5nt)

[Ví dụ dạng categorical data 5](#_y4wvje3jlonk)

[Phát biểu thuật toán kNN; data & results 5](#_cu3q7j7g6q6g)

[Decision boundary của kNN được tính như nào? 5](#_7nt5s2sx1msq)

[Choosing k ra sao? 6](#_yp3veps4e2s8)

[Input normalization có ý nghĩa gì? Tại sao? 6](#_cfti2dl10xb)

[Có những cách normalization như thế nào? 6](#_o32pt81pg3q5)

[Decision Trees hoạt động như thế nào? 6](#_lfssfbdp4ioc)

[Mô tả thuật toán học Regression Tree 6](#_cw5soov1oviw)

[Stopping criterion trong decision tree là như nào? 6](#_9y4xeg5i0n2e)

[Depth của Decision tree được chọn như thế nào? 6](#_sgvbld68nge7)

[**Chapter 3 6**](#_cso1jv349x8g)

[Phát biểu linear regression model 6](#_hol8en8h2n8)

[Loss function là gì? 6](#_p00kd2cv1rzd)

[Loss function khác cost function như nào? 6](#_pk1cb0i0pbvp)

[Operator argmin\_theta nghĩa là gì? 6](#_vyvqirmlihms)

[Least squares là gì? 6](#_qxabtmc6mwgi)

[Solution của LR có dạng như thế nào? 6](#_ycuertc3x5h2)

[Normal equations là gì? 6](#_foqqrxivaenj)

[Quan hệ giữa least squares method và maximum likelihood? 6](#_15irsdzt8jp)

[Mô tả Maximum likelihood solution 7](#_qrfimrood1tg)

[Logistic regression để làm gì? 7](#_11hoyyansfzs)

[Logistic function có dạng như nào? 7](#_rhx07l203vk6)

[Binary cross-entropy loss có dạng như nào? 7](#_8w9tn8ibpk94)

[Negative log-likelihood cost function có dạng như nào? 7](#_8e40tgnzls0n)

[cross-entropy loss là gì? 7](#_yrzv7a25jl22)

[Thuật toán logistic regression cho nhiều hơn 2 classes? 7](#_c1oby7jr3d90)

[Polynomial regression được thực hiện như nào? 7](#_nkf99peu9t2)

[Regularization để làm gì? 7](#_jmkfzul9jqqd)

[Tên khác của linear regression với L2 regularization 7](#_y3kzeltngs71)

[Generalized linear models dùng để làm gì? 7](#_x78mqb7j3qiw)

[Link function có tác dụng gì? 7](#_ux5ujf89qspj)

[Một GLM bao gồm những gì? 7](#_hwkwjclxyhwg)

[**Chapter 4 7**](#_wdiwk7yfuuku)

[Train, val, test có ý nghĩa như nào? 7](#_ehnrteabplrh)

[Loss và metrics khác gì nhau? 7](#_hbd0gxbr7x4m)

[Ví dụ các metrics dùng cho regression 8](#_wr07x486y761)

[Ví dụ cá metrics dùng cho classification 8](#_wgob8iegnoc0)

[Bias là gì? 8](#_2p8t3iz6lcr1)

[Vaiance là gì? 8](#_9vl9bp1cz8qg)

[E\_new để đo cái gì? 8](#_c5snutnbi2t9)

[Ta có thể decompose E\_new ra thành gì? 8](#_2a4m4rpuvn4o)

[Irreducible error là gì? 8](#_7bs3lu6dm967)

[Overfitting là gì? 8](#_9c9rieddxz5u)

[Underfitting là gì? 8](#_2f1khaeytjrq)

[Underfitting và overfitting liên quan như nào tới model complexity? 8](#_oswdg4duboc7)

[Training size của data ảnh hưởng bias và variance như nào? 8](#_izkp6tv99f7r)

[Bias-variance tradeoff là gì? 8](#_tiq15t3sqjj7)

[Generalization gap là gì? 8](#_q4wgbzg32cnt)

[Confusion matrix là gì? 8](#_9da72opy0ced)

[ROC curve là gì? 8](#_hs5k10jktz3s)

[Precision và Accuracy khác gì nhau? 8](#_y8od9j9v30xe)

[ROC AUC = 0.7 có ý nghĩa gì? 9](#_o3swjksk4s4p)

[F1 score có ý nghĩa gì? 9](#_9z80zfwm9rn3)

[Precision và Recall thì cái gì quan trọng hơn? 9](#_rhxaye7r2puc)

[**Chapter 5 9**](#_sc0t299u637h)

[Ví dụ của parametric models 9](#_z6xkzp6ydbfo)

[Early stopping là gì? 9](#_dtz33kb2ynzd)

[Regularization để làm gì? 9](#_7e88n9m7s10i)

[Robustness của 1 model có ý nghĩa gì? 9](#_76wtiytm208u)

[Ví dụ 3 loss functions có thể dùng cho bài toán classification 9](#_v2smvu13cjwq)

[Ví dụ 3 loss functions có thể dùng cho regression 9](#_8k3rba55ortz)

[Ví dụ loss function dùng cho multiclass classification 9](#_vdr4fabf4cjh)

[Kể tên các loại regularization 9](#_z581no89o38t)

[Gradient descent hoạt động như thế nào? 9](#_eh6xrmnidzb7)

[Stochastic gradient descent là gì? 9](#_bwctag7q319u)

[Parameter & hyperparameter optimization khác gì nhau? 9](#_659vgybf4omo)

[Learning rate có ý nghĩa như nào? 9](#_g0vx01m0r8li)

[Có thể thay đổi learning rate trong quá trình training không? 10](#_svb1epynqo25)

[Các phương pháp tương tự gradient descent? 10](#_r70lkhkcrtjp)

[**Chapter 7 10**](#_gs0t6y8aca47)

[Ensemble methods dùng để làm gì? Có mấy loại? 10](#_u4fmeyf9hcxg)

[Bagging là gì? 10](#_x4ub9h1vq2o)

[Boostrap có ý nghĩa gì? 10](#_94l9wz7mwcub)

[Boosting để làm gì? 10](#_zb561bccx13l)

[Random forest hoạt động như nào? 10](#_au9mph5w0ayk)

[Bagging làm giảm hay tăng variance? 10](#_fdj7x2o8kq7f)

[Bagging có làm thay đổi bias không? 10](#_x3cws2fa8d4c)

[Ensemble có kiến model bị overfitting không? 10](#_7zdzs59ht6q9)

[Mô tả thuật toán Adaboost 10](#_c1f08no2f5dp)

[Mô tả thuật toán Gradient boosting 10](#_4t1uklk3eee5)

[**Chapter 8 10**](#_816y1euwym8y)

[Kernel trick là gì? 10](#_ypnenxwnwtdk)

[Kernel để làm gì? 10](#_682q5dfj2biz)

[Support vectors trong SVM có ý nghĩa gì? 11](#_kiaz2ksvjmfz)

[Dual parameters có ý nghĩa gì? 11](#_fxlu5psvr9xd)

[Có điều gì đặc biệt khi sử dụng epsilon-insensitive loss function kết hợp với kernel method trong regression? 11](#_67ydea3o4vit)

[Có điều gì đặc biệt khi sử dụng hinge loss function kết hợp với kernel method trong classification? 11](#_6j2s3plkn1sq)

[Liệt kê 3 ý nghĩa của kernel trong ML 11](#_4op555xo3jgc)

[Điều gì ảnh hưởng đến việc chọn kernels? 11](#_a2sdaf1ms7ut)

[Ví dụ 3 kernels 11](#_p729t6s08kcj)

[**Chapter 9 11**](#_tdv7mtt8e5bm)

[Điều khác biệt của Bayesian approach với parametric model trong ML là gì? 11](#_m6byt0cevz2x)

[Nêu ý nghĩa của từng loại xác suất trong Bayesian approach: p(theta|y) , p(y\_star|y), p(theta), p(y), p(y|theta) 11](#_bxplxoinj7z2)

[Mô tả mô hình linear regression sử dụng Bayesian approach (i.e., Bayesian linear regression) 11](#_5xp2cmfa6dsl)

[Maximum a posterior (MAP) point có ý nghĩa gì và được xác định như nào? 11](#_5qdqy9pwn8fe)

[Gaussian process là gì? 11](#_nl0grq9gb0lr)

[**Chapter 10 12**](#_qafvvqvya96z)

[Phân biệt discriminative models và generative models 12](#_pnmio421kp97)

[Generative model sử dụng phân bố nào? 12](#_xyi2sevo4cdp)

[Gaussian Mixture Model dùng để làm gì? 12](#_cigp06rn2r4d)

[Giả thuyết cơ bản của GMM là gì? 12](#_20pvdvel3dth)

[Mô tả cách học 1 GMM 12](#_tqnhbhu93qa9)

[Mô tả cách predict với GMM 12](#_17pz2w6qr0pi)

[Mô tả cách học semi-supervised learning với GMM 12](#_1n1b73j3x98m)

[Mô tả cách dùng GMM cho Cluster analysis 12](#_xzt8yz5qyy1s)

[Mô tả thuật toán k-means 12](#_yfo1ywk0c1ys)

[**Chapter 11 12**](#_l2z35m8gohne)

[User aspects trong machine learning bao gồm những gì? 12](#_x78m4yz8kobx)

[Làm sao để tạo một baseline để đo performance level? 12](#_djtep8sin9z7)

[Làm thế nào để biết model có đáng tin hay không? 12](#_2rpzkzxec3u)

[**Chapter 12 12**](#_uxejpgs9h993)

[Các vấn đề đạo đức trong ML bao gồm những gì? 12](#_31j46tbb2nn1)

[Ethics through awareness để chỉ điều gì? 13](#_opti4mrzq5k)

[Hãy nêu 3 ví dụ các vấn đề đạo đức khi ứng dụng ML 13](#_dt0nxbv98kcf)

[Có thể đạt được fairness sử dụng metrics nào? 13](#_67i4xee07ey8)

[Có các kỹ thuật nào để phát hiện bias trong ML model? 13](#_ov01evp0j9o9)

[Mô tả 1 hype trong ML và phân tích các khía cạnh đạo đức 13](#_adx47lqie7bv)

# 

# Chapter 1

## Học máy là gì?

* Là học, lý luận, quyết định dựa trên data và được tiến hành bởi chương trình máy tính sử dụng data nhằm lấy ra thông tin hữu ích, đưa ra dự đoán, gợi ý về hành động hoặc quyết định hợp lý

## Học từ data (learnt from data) là gì?

* Học từ dữ liệu là quá trình hệ thống máy tính tự động hóa việc cải thiện hiểu biết của nó từ những dữ liệu quan sát được.

## Model là gì?

* Mô hình là một biểu diễn đầy đủ của dữ liệu, dưới dạng toán học chính xác, nó nắm bắt được các đặt điểm quan trọng của hiện tượng mà chúng ta đang nghiên cứu. Một mô hình bao gồm các tham số được điều chỉnh trong quá trình đào tạo dựa trên tập dữ liệu đầu vào.

## Learning algorithm là gì?

* Thuật toán học là một quy trình toán học cụ thể được sử dụng để huấn luyện mô hình trên dữ liệu. Thuật toán sẽ điều chỉnh các tham số của mô hình sao cho mô hình có khả năng quyết định và dự đoán chính xác dựa trên dữ liệu mới.

## 3 phần quan trọng nhất của học máy là gì?

* The data
* The mathematical model
* The learning algorithm

# Chapter 2

## Supervised, Unsupervised, Semi-supervised

* Supervised: dữ liệu dùng để huấn luyện mô hình có chứa cả nhãn/đầu ra dữ liệu mong muốn
* Unsupervised: dữ liệu dùng để huấn luyện mô hình không có nhãn/đầu ra dữ liệu mong muốn
* Semi-supervised: sự kết hợp của 2 phương pháp trên (dữ liệu bị thiếu nhãn)

## Training data được tạo ra như thế nào?

Sometimes it’s just recording X and Y but sometimes we need an expert in that domain to label the data. For ex : detecting cancer from a X-ray image

Đôi khi chỉ ghi X và Y nhưng đôi khi chúng tôi cần một chuyên gia trong lĩnh vực đó để gắn nhãn dữ liệu. Ví dụ: phát hiện ung thư từ ảnh X-quang

## Regression vs Classification

Class ⇒ output is categorical, Regress ⇒ output is numerical (xem ở dưới để biết thêm)

## Numerical and categorical variables

Numerical

* Là dữ liệu liên tục như số thực hoặc có thể rời rạc như số nguyên
* Có thứ tự tự nhiên (bạn có thể biết cái nào lớn hơn, có thể so sánh được)
* Ví dụ: tốc độ của ô tô, giá nhà, trọng lượng của một người

Categorical

* Luôn rời rạc
* Thiếu trật tự tự nhiên
* Ví dụ: người đàn ông này có lương > 5k không, 1 = Thụy Điển 2 = Na Uy 3 = Phần Lan,….

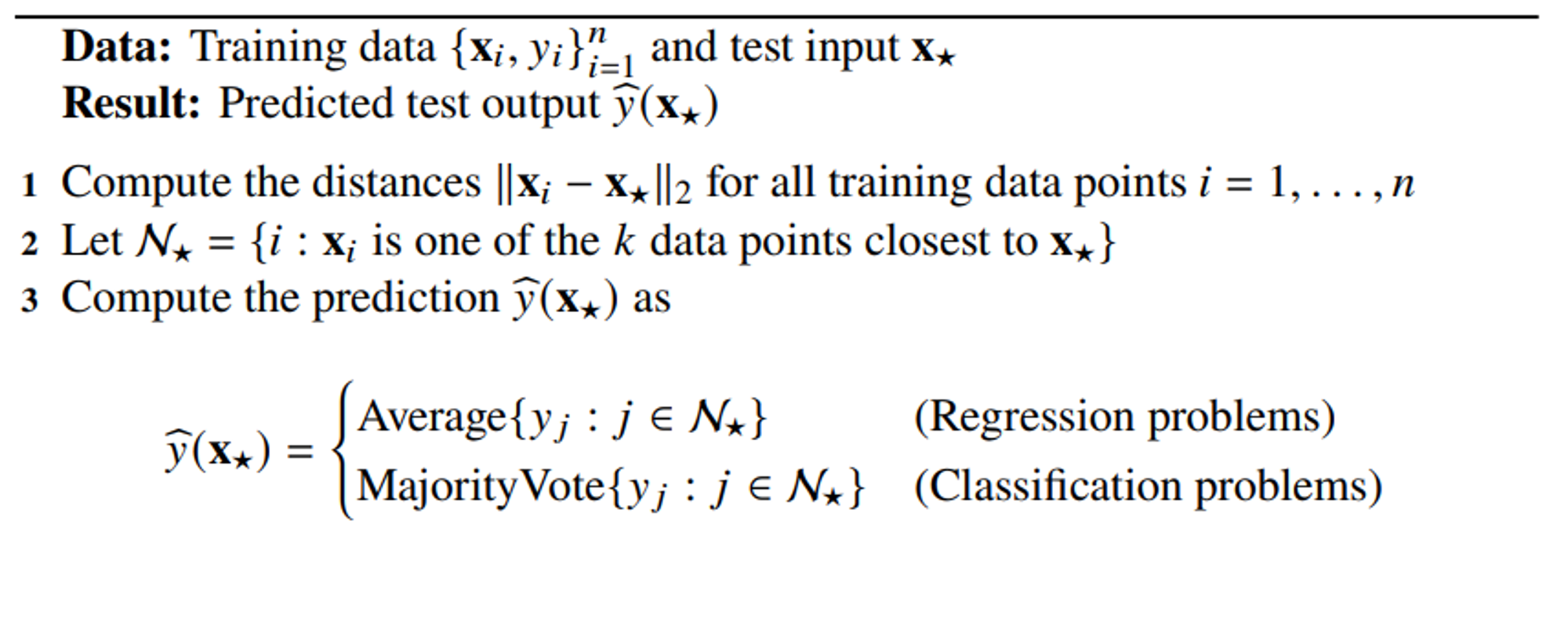
## Ví dụ dạng numerical data

(above)

## Ví dụ dạng categorical data

(above)

## Phát biểu thuật toán kNN; data & results



* Dữ liệu: Toàn bộ tập dữ liệu có dạng X, y (X = features, y = prediction)
* Thuật toán
* Tính khoảng cách (usually norm\_2) giữa điểm x\* cần dự đoán với toàn bộ điểm x thuộc tập dữ liệu
* Chọn ra k điểm gần nhất với x\*
* Tính prediction:
* Avg với regression
* Majority vote với classification

## Decision boundary của kNN được tính như nào?

* kNN không tính trực tiếp decision boundary nhưng ta có thể infer được decision boundary của nó



* Cách làm: [knn-bounds+soln (mit.edu)](http://web.mit.edu/6.034/wwwbob/knn-bounds+soln.pdf)

## Choosing k ra sao?

* Với k quá nhỏ (như 1), chúng ta khớp dữ liệu huấn luyện cực kỳ tốt nhưng mô hình lại quá khớp
* Với k quá lớn chúng ta sẽ loại bỏ tất cả các mẫu có thể quan sát được
* Chọn k theo giá trị chéo (k bằng cách nào đó rơi vào khoảng 10-35, không ai biết chắc chắn)

## Input normalization có ý nghĩa gì? Tại sao?

* Chuẩn hóa đầu vào là làm cho tất cả tính năng dữ liệu có cùng phạm vi.
* Khi một phạm vi quá lớn, phương pháp dựa trên khoảng cách của chúng tôi như KNN sẽ đánh giá quá cao tính năng có phạm vi lớn, bỏ qua tính năng khác. Ta muốn tất cả các tính năng đều có tầm quan trọng như nhau

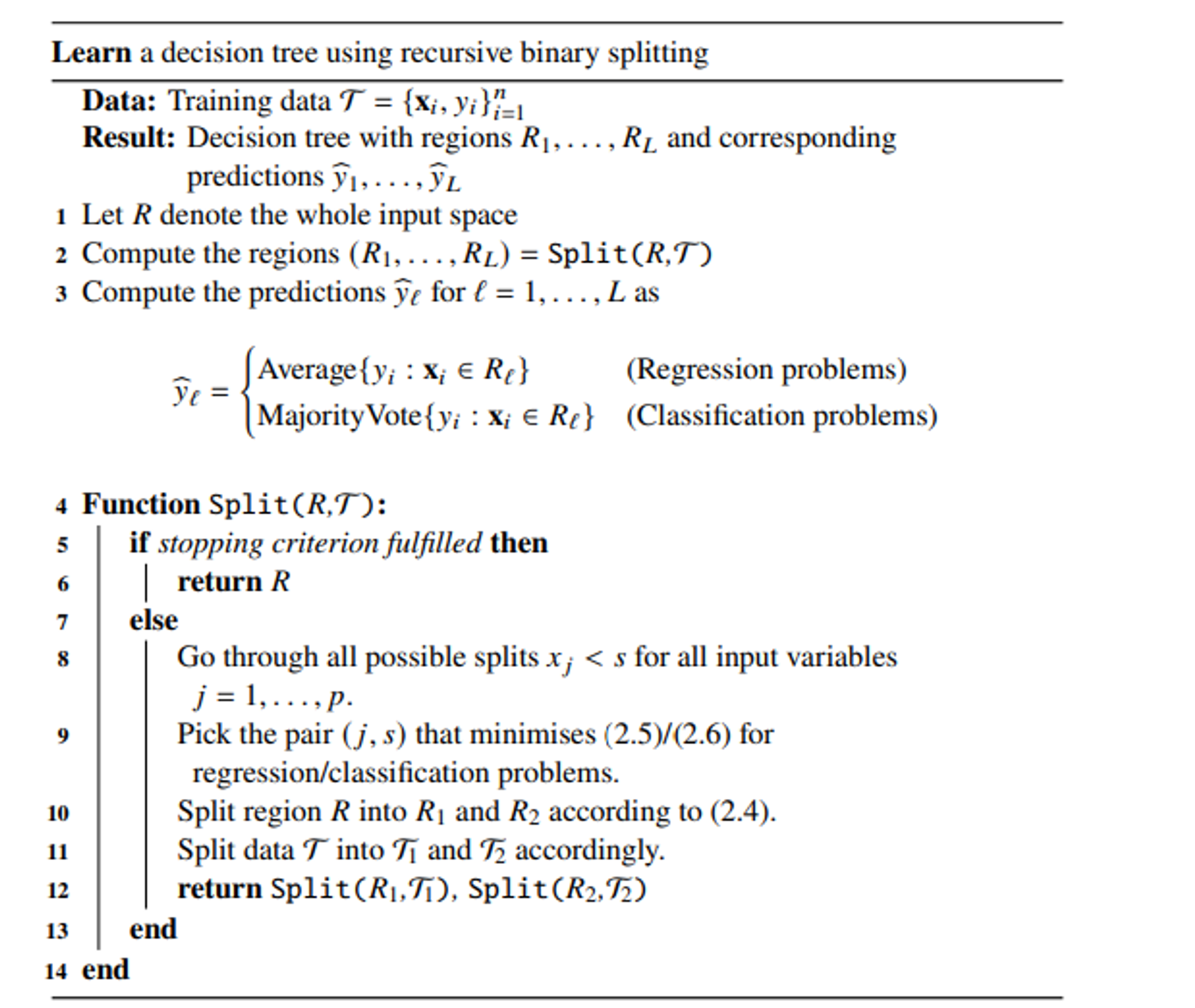
## Có những cách normalization như thế nào?

* [0, 1] range:
* Z-score → mean = 0 and std = 1:

## Decision Trees hoạt động như thế nào?

* Cây quyết định (Decision tree) là một tập hợp các quy tắc đã học
* Cây quyết định (Decision tree) phân chia hiệu quả không gian đầu vào thành nhiều vùng riêng biệt và trong mỗi vùng, một giá trị không đổi được sử dụng để dự đoán

## Mô tả thuật toán học Regression Tree



* Dữ liệu: features X và output y tương ứng
* Kết quả: một cây gồm L lá R\_1 -> R\_L
* Cách chia R\_1 -> R\_L
* Duyệt toàn bộ các cách chia cây với điều kiện x\_j < s khác nhau, và chọn cặp (j, s) sao cho chúng thỏa mãn độ giảm Gini index/entropy lớn nhất

Gini index: weighted sum of

Entropy: weighted sum of

## Stopping criterion trong decision tree là như nào?

* Một phương pháp ngăn chặn cây phát triển hoàn toàn (mô hình kết quả khá giống với KNN với k = 1)
* VD:
* Giới hạn độ sâu tối đa
* Giới hạn số lượng điểm dữ liệu huấn luyện tối thiểu được liên kết với mỗi nút lá

## Depth của Decision tree được chọn như thế nào?

* Nếu chúng ta để cây trở nên quá sâu, mô hình thu được khá giống với KNN với k = 1 ⇒ overfit
* Nếu chúng ta để cây không đủ sâu, nó sẽ không nắm bắt được bất kỳ sự kết hợp có ý nghĩa nào giữa các đặc điểm
* Đó là sự đánh đổi giữa flexibility và rigidity nên chúng ta có thể chọn K cho KNN, k-fold (bias-variance tradeoff basically)

# Chapter 3

## Phát biểu linear regression model

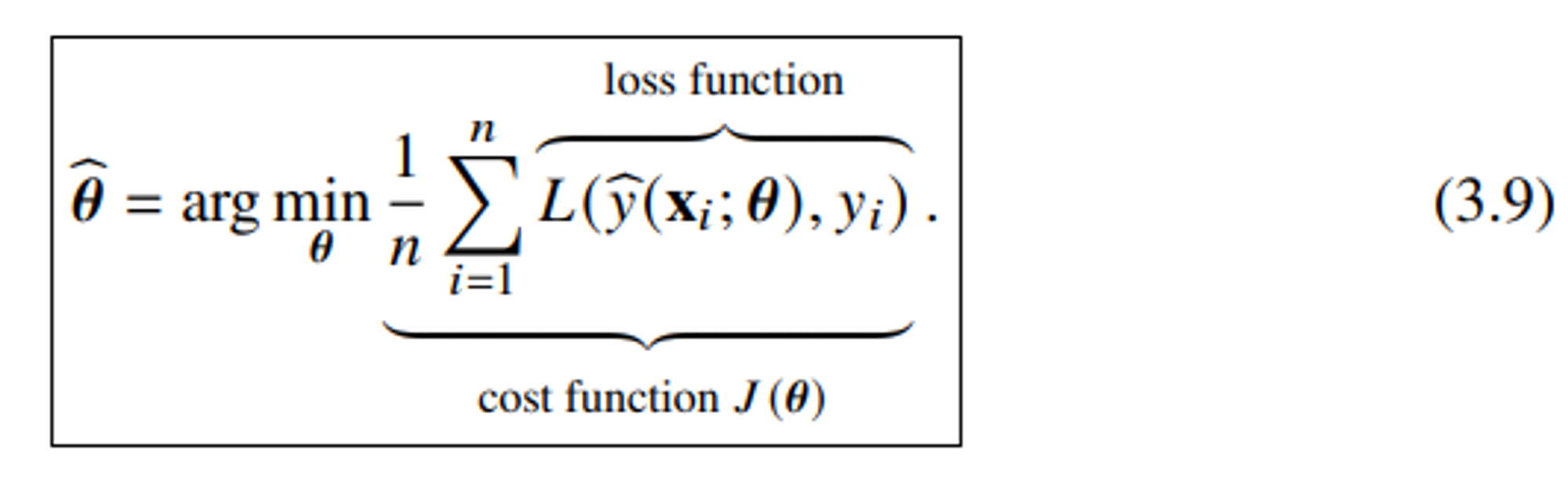
* Mô hình hồi quy tuyến tính giả sử output y có thể được biểu diễn dưới một affine combination ( basically)
* = parameters, = y-intercept, intercept/offset term
* Noise không được capture trong model, và chúng được giả sử có normal distribution:
* Công thức:

## Loss function là gì?

Đo sự “gần” của prediction so với ground truth, được ký hiệu

Nếu model fit với data tốt thì và loss function nhỏ

## Loss function khác cost function như nào?



loss function is define for a prediction, cost function is define for a model (check this for me pls)

## Operator argmin\_theta nghĩa là gì?

The value theta where the function achieve its minimum

## Least squares là gì?

Một cách định nghĩa cost function của Linear regression:

## Solution của LR có dạng như thế nào?

→ Solution = nghiệm normal equation =

## Normal equations là gì?

(differential of argmin function above)

## Quan hệ giữa least squares method và maximum likelihood?

Maximum likelihood estimation

* Mục tiêu: tìm sao cho:
* Giả sử n data points là iid →
* Giả sử là normal distribution:

* As log is monotonically increasing, we have

which is the definition of least square method

## Mô tả Maximum likelihood solution

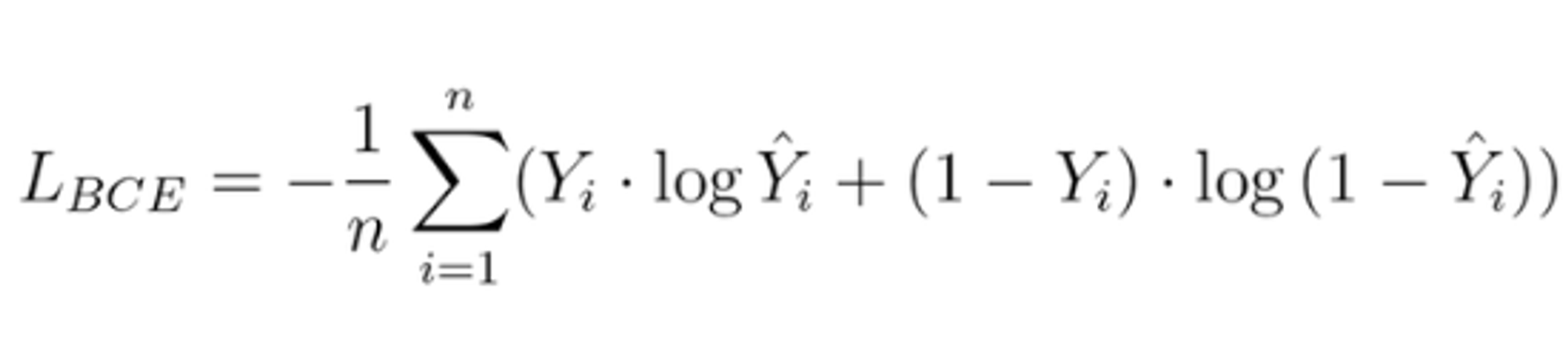
(above)

## Logistic regression để làm gì?

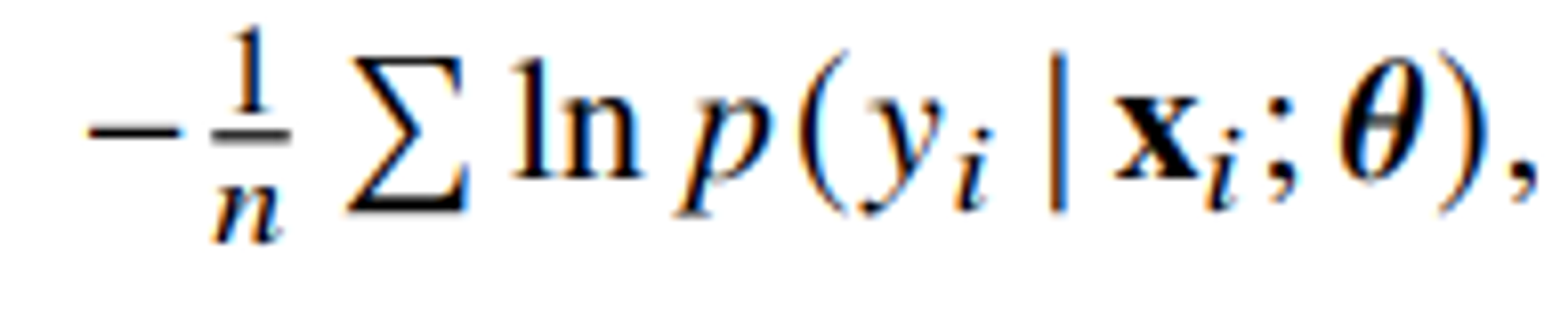
* Logistic regression: Linear regression + logistic function
* Dùng cho các bài toán classification

## Logistic function có dạng như nào?

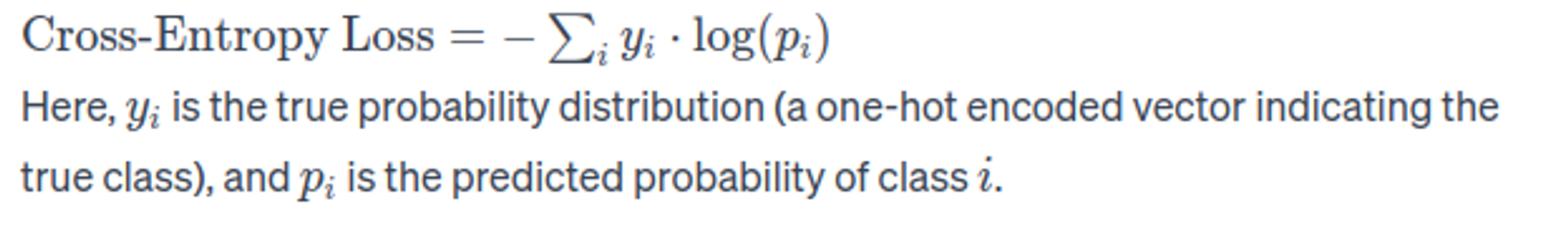
## Binary cross-entropy loss có dạng như nào?



## Negative log-likelihood cost function có dạng như nào?



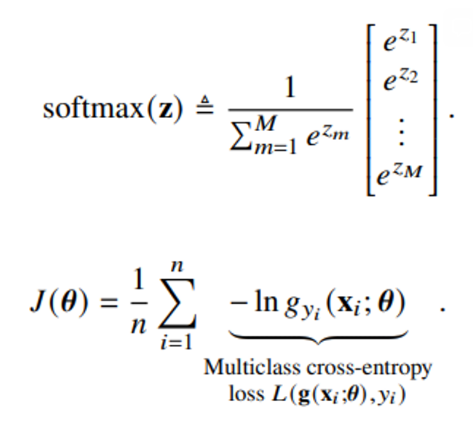
## cross-entropy loss là gì?



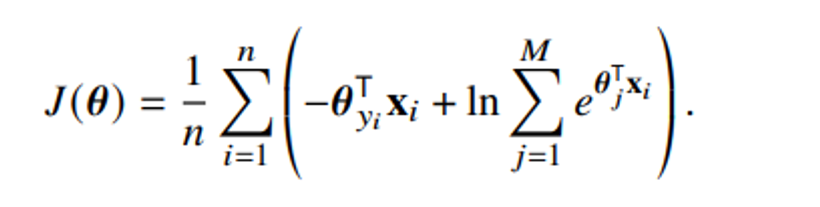
## Thuật toán logistic regression cho nhiều hơn 2 classes?

C1: one-vs-all

C2: softmax



Optimization problem:



## Polynomial regression được thực hiện như nào?

Input X với chiều là 1 → tăng chiều lên thành X^2, X^3, etc.

## Regularization để làm gì?

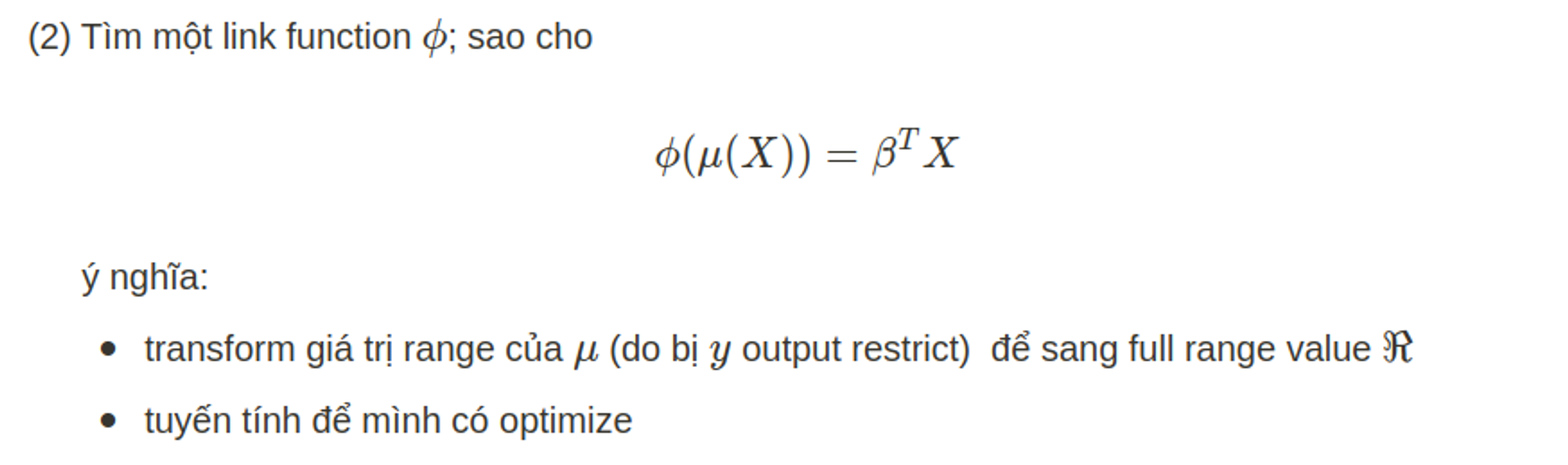
* Intuition: dùng theta nhỏ cho dimension cao thay vì theta to để giảm overfitting
* Nói cách khác, khi polynomial regression, regularization giúp giảm overfitting

## Tên khác của linear regression với L2 regularization

* Ridge Regression

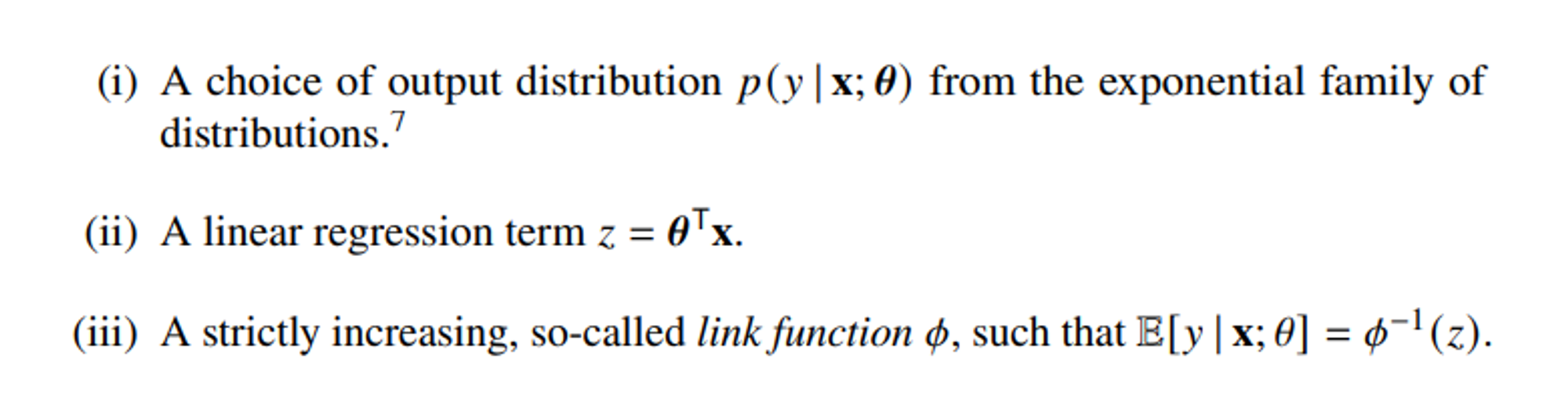
## Generalized linear models dùng để làm gì?

## Link function có tác dụng gì?



* Mặt khác, link function cho chúng ta biết về giá trị trung bình của X và theta ⇒ biết phân bố chính xác mà chúng ta đang tính toán ⇒ so sánh tổn thất với phân bố dữ liệu

## Một GLM bao gồm những gì?



# Chapter 4

## Train, val, test có ý nghĩa như nào?

* Training data: dữ liệu huấn luyện
* Validation data: dự đoán test error (do E\_train != E\_new, nhưng E\_val ~ E\_new)

## Loss và metrics khác gì nhau?

* Metrics: là các giá trị số định lượng mức độ phù hợp của mô hình với dữ liệu thực tế hoặc đầu ra mong muốn. Sau khi huấn luyện xong, các metrics sẽ được sử dụng đến đánh giá mô hình
* Loss là các hàm toán học tính toán sự khác biệt hoặc sai số giữa dự đoán của mô hình và giá trị thực. Chúng được sử dụng để tối ưu mô hình: mô hình sẽ tối thiểu hóa loss function (của tập huấn luyện) trong lúc huấn luyện

## Ví dụ các metrics dùng cho regression

* MAE, MSE, RMSE = sqrt(MSE)
* R^2 score,

## Ví dụ cá metrics dùng cho classification

* F1, Fbeta score
* AUC-ROC, AUC-PR
* Accuracy

## Bias là gì?

* Bias là lỗi được tạo ra bằng cách xấp xỉ một vấn đề trong thế giới thực bằng cách sử dụng một mô hình đơn giản hóa. Nó đo lường mức độ khác biệt của các dự đoán của mô hình so với giá trị thực.
* Bias gắn liền với các giả định do mô hình đưa ra và tính đơn giản trong cách biểu diễn các mẫu cơ bản trong dữ liệu.
* Bias cao có thể do mô hình không nắm bắt được độ phức tạp của dữ liệu, dẫn đến hiệu suất kém trên cả tập huấn luyện và tập kiểm tra.
* Ví dụ: Giá nhà tỉ lệ thuận với bình phương diện tích nhà, thì bias cao có thể là do ta chỉ dùng hồi quy tuyến tính, dẫn đến bias cao. Ta có thể giảm bias bằng cách sử dụng một mô hình hồi quy “đa tuyến” (polynomial regression)

## Variance là gì?

* Phương sai đo lường độ nhạy của mô hình đối với những biến động nhỏ hoặc nhiễu trong dữ liệu huấn luyện. Nó thể hiện mức độ thay đổi trong dự đoán của mô hình nếu nó được huấn luyện trên một tập hợp con dữ liệu khác.
* Mô hình có phương sai cao quá linh hoạt và thường capture cả noise trong dữ liệu huấn luyện thay vì các mẫu cơ bản. Điều này có thể dẫn đến tình trạng overfitting, trong đó mô hình hoạt động tốt trên tập huấn luyện nhưng lại hoạt động kém trên dữ liệu mới, chưa được nhìn thấy.
* Nguyên nhân phổ biến dẫn đến phương sai cao bao gồm việc sử dụng một mô hình rất phức tạp, có quá nhiều tính năng hoặc không có đủ dữ liệu để huấn luyện mô hình một cách hiệu quả.

## E\_new để đo cái gì?

Là giá trị của hàm mất mát đối với tập dữ liệu kiểm tra (test data)

## Ta có thể decompose E\_new ra thành gì?

Cách 1: Training error-Generalisation gap

Cách 2: Bias-Variance Decomposition

E\_new = Bias^2 + Variance + Irreducible Error

## Irreducible error là gì?

Là lỗi của mô hình mà không thể cải thiện bằng bất cứ phương pháp nào, hay lỗi của mô hình nằm ngoài tầm kiểm soát của mô hình. Ví dụ, sự sai sót trong đo đạc có thể là một nguyên nhân của irreducible error

## Overfitting là gì?

Overfitting là tình huống trong đó một mô hình quá phức tạp dẫn đến mô hình hoạt động tốt, capture cả noise, trên tập huấn luyện nhưng lại hoạt động kém trên dữ liệu mới, chưa được nhìn thấy.

## Underfitting là gì?

Underfitting là một tình huống trong học máy trong đó một mô hình quá đơn giản để nắm bắt được các mẫu cơ bản trong dữ liệu. Kết quả là mô hình hoạt động kém không chỉ trên dữ liệu huấn luyện mà còn trên dữ liệu mới, chưa được nhìn thấy.

## Underfitting và overfitting liên quan như nào tới model complexity?

Underfitting: model quá đơn giản

Overfitting: model quá phức tạp

## Training size của data ảnh hưởng bias và variance như nào?

Training size không ảnh hưởng bias, nhưng tăng training size có thể giảm variance, nếu dữ liệu được thêm vào có thêm sự đa dạng (về mặt distribution, về khu vực được lấy, etc.)

## Bias-variance tradeoff là gì?

Khi model complexity tăng thì khả năng cao là bias sẽ giảm, nhưng variance lại tăng

Khi model complexity giảm thì khả năng là bias sẽ tăng, nhưng variance lại giảm.

Như vậy, bias-variance tradeoff là sự lựa chọn model complexity cho hợp lý giúp cân bằng bias và variance trong học máy

## Generalization gap là gì?

Là khoảng cách giữa E\_train và E\_new:

Mô hình càng phức tạp, generalisation gap càng cao (high variance)

## Confusion matrix là gì?

Confusion matrix thể hiện sự đánh nhãn đúng sai với từng nhóm

Với bài toán phân lớp nhị phân (binary classification,) confusion matrix là ma trận gồm 4 phần tử TP, FP, TN, FN

## ROC curve là gì?

ROC: hàm xấp xỉ giá trị của TPR, FPR với các threshold khác nhau

## Precision và Accuracy khác gì nhau?

Precision:

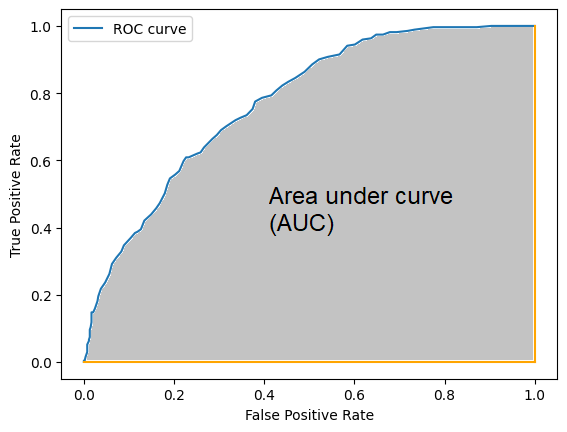
* Tỉ lệ số điểm true positive trong số những điểm được phân loại là positive (TP + FP).

Recall:

* Tỉ lệ số điểm true positive trong số những điểm thực sự là positive (TP + FN).

## ROC AUC = 0.7 có ý nghĩa gì?

AUC-ROC: diện tích dưới hàm ROC đề cập bên trên (hình minh họa phía dưới)

* 

## F1 score có ý nghĩa gì?

* Là con số kết hợp cả Precision và Recall vào là

## Precision và Recall thì cái gì quan trọng hơn?

Tùy vào bối cảnh bài toán mà Precision hoặc Recall có thể quan trọng hơn

Ta lấy ví dụ đầu tiên là một bài toán dự đoán 1 người có ung thư hay không. Ở đây thì việc dự đoán sai người không ung thư là người ung thư (FP) sẽ tốt hơn là việc bỏ sót người ung thư (FN.) Vi vậy, chỉ số Recall sẽ quan trọng hơn Precision ở trường hợp này.

Ở một ví dụ khác, với bài toán phát hiện spam email thì việc bỏ sót một số email spam (FN) sẽ tốt hơn là việc loại bỏ/đánh dấu những email quan trọng (FP.) Từ đó, Precision sẽ quan trọng hơn Recall ở bài toán này.

Để đánh giá chung Precision và Recall khi một trong 2 chỉ số trên quan trọng hơn chỉ số còn lại, ta có thể dùng F\_beta score. Với beta < 1 thì Precision quan trọng hơn Recall, và beta > 1 thì Recall quan trọng hơn Precision. Với beta = 1 thì đó là F1 score như đã đề cập ở trên.

# Chapter 5

## Ví dụ của parametric models

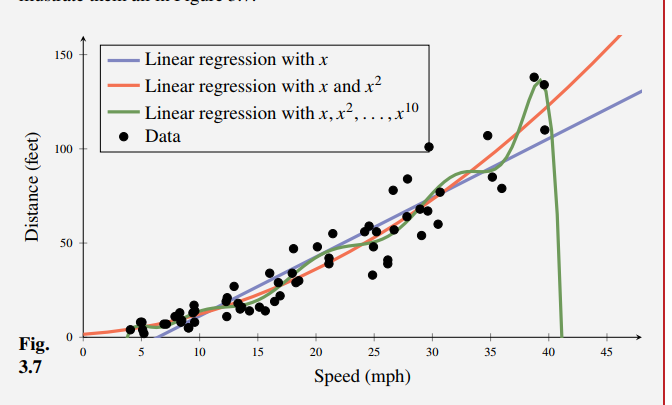
Linear regression and logistic regression

## Early stopping là gì?

dừng thuật toán tối ưu hóa sớm

## Regularization để làm gì?

Với các mô hình hồi quy “đa tuyến” (polynomial regression,) khi chúng ta cố gắng “fit” mô hình vào điểm dữ liệu, tình trạng dưới đây có thể xảy ra



Đây là tình trạng overfitting, hay mô hình quá tập trung vào “fit” vào các điểm dữ liệu trong tập huấn luyện mà quên mất còn những dữ liệu ở tập kiểm tra (test)

Để giảm tình trạng này, regularisation được sử dụng. Cụ thể, một phần tử lambda sẽ được thêm vào loss function

## Robustness của 1 model có ý nghĩa gì?

robust = less prone to outlier

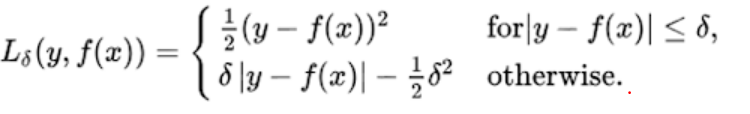
## Ví dụ 3 loss functions có thể dùng cho bài toán classification

* Binary Cross entropy (for binary classification)
* Logistic Loss (for n-classification, n >= 2) - Binary Cross entropy is for n=2
* Square Hinge loss
* Hinge loss: dùng cho SVM - binary classification

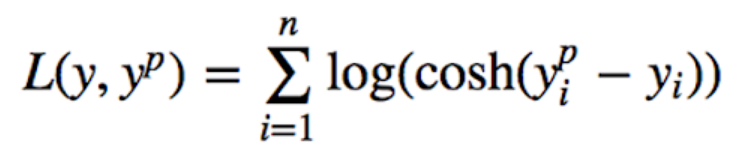
với y = 1, -1

## Ví dụ 3 loss functions có thể dùng cho regression

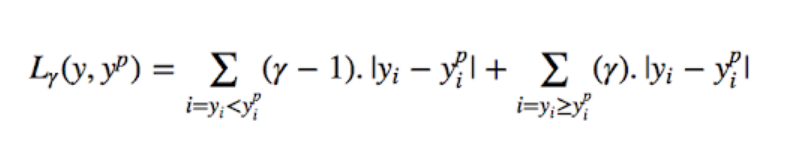
* L1 loss - mean abs error
* L2 loss - squared error
* Huber loss: L2 loss khi error nhỏ, L1 loss khi error to



* Log-cosh loss:



* Quantile loss:



## Ví dụ loss function dùng cho multiclass classification

## Kể tên các loại regularization

* Ridge Regression (L2 norm Regularization)
* Lasso Regression (L1 norm Regularization)

## Gradient descent hoạt động như thế nào?

## Stochastic gradient descent là gì?

## Parameter & hyperparameter optimization khác gì nhau?

## Learning rate có ý nghĩa như nào?

## Có thể thay đổi learning rate trong quá trình training không?

## Các phương pháp tương tự gradient descent?

# Chapter 7

## Ensemble methods dùng để làm gì? Có mấy loại?

## Bagging là gì?

## Boostrap có ý nghĩa gì?

## Boosting để làm gì?

## Random forest hoạt động như nào?

## Bagging làm giảm hay tăng variance?

## Bagging có làm thay đổi bias không?

## Ensemble có kiến model bị overfitting không?

## Mô tả thuật toán Adaboost

## Mô tả thuật toán Gradient boosting

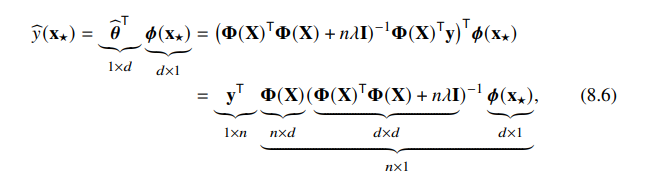
# Chapter 8

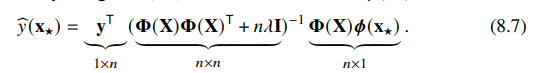
## Kernel trick là gì?

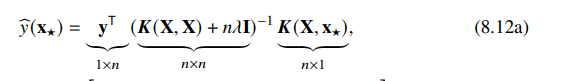
* Là một kỹ thuật mà ta có thể chọn thẳng k(x, x’) thay vì 𝝓(x) nếu inputs của models chỉ ở dưới dạng 𝝓(x)^T.𝝓(x).

## Kernel để làm gì?

* Kernel là bất kỳ hàm số nào lấy x và x’ từ cùng một không gian và trả về một scalar



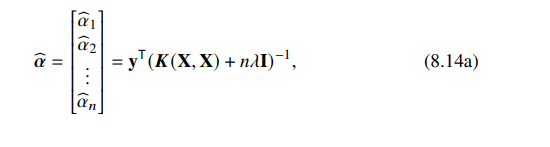


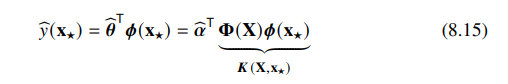


=> Ta phải chọn kernel sao cho 𝑲(X, X) + 𝑛𝜆I khả nghịch

## Support vectors trong SVM có ý nghĩa gì?

## Dual parameters có ý nghĩa gì?

* Với công thức 8.12a) mỗi lần predict ta lại phải tính phần inverse -> điều này là không cần thiết vì phần inverse không phụ thuộc vào input vì thế ta có dual parameter:
* 
* Nhờ vậy ta viết lại được bài toán dưới dạng



=> Ý nghĩa:

* Giúp cài đặt dễ dàng hơn
* Ta chuyển từ việc chọn tham số theta (với số chiều là d -> do ta chọn) -> dual parameter phụ thuộc vào n (số lượng điểm dữ liệu)

## Có điều gì đặc biệt khi sử dụng epsilon-insensitive loss function kết hợp với kernel method trong regression?

## Có điều gì đặc biệt khi sử dụng hinge loss function kết hợp với kernel method trong classification?

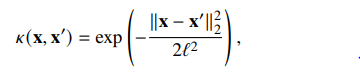
## Liệt kê 3 ý nghĩa của kernel trong ML

## Điều gì ảnh hưởng đến việc chọn kernels?

## Ví dụ 3 kernels







(từ trên xuống dưới - linear kernel, polynomial kernel, squared exponential kernel)

# Chapter 9

## Điều khác biệt của Bayesian approach với parametric model trong ML là gì?

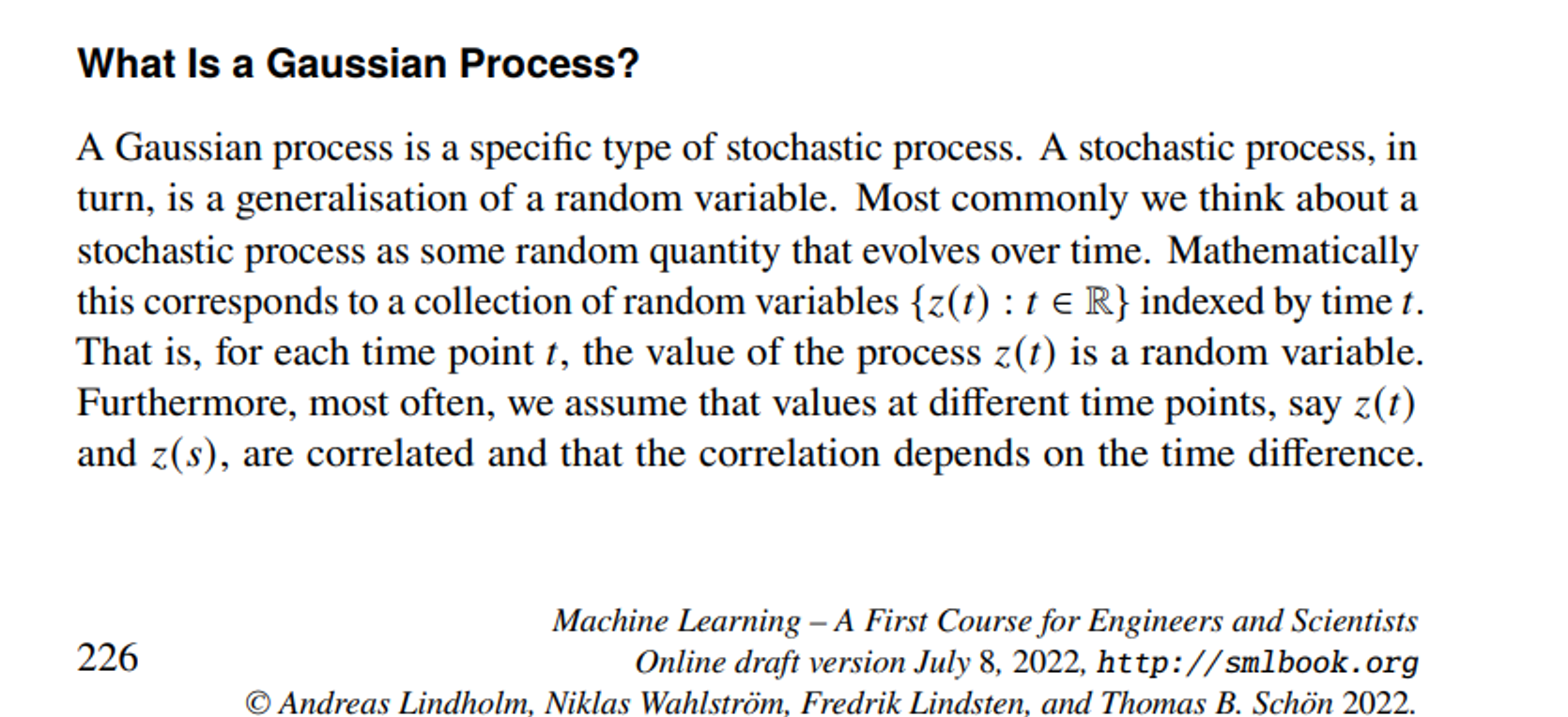
## Nêu ý nghĩa của từng loại xác suất trong Bayesian approach: p(theta|y) , p(y\_star|y), p(theta), p(y), p(y|theta)

## Mô tả mô hình linear regression sử dụng Bayesian approach (i.e., Bayesian linear regression)

## Maximum a posterior (MAP) point có ý nghĩa gì và được xác định như nào?

## Gaussian process là gì?

* Đó là gaussian approach cho mô hình phi tham số
* Nói tóm lại, đó là gaussian approach với kernel Ridge regression



# Chapter 10

## Phân biệt discriminative models và generative models

* Discriminative models: (conditional model) -> học từ data để dự đoán output theo điều kiện từ input, estimate **p(Y|X)** 
  + Aim: maximize the separation between class in dataset to perform classification or regression
* Generative models: Mô tả phân bố xác suất của cả inputs lẫn outputs, estimate **p(X, y)** 
  + Is the family of model that create new data point
  + **p(X, y) => p(Y|X)**
  + **p(X, y) = p(Y|X).p(X)**

## Generative model sử dụng phân bố nào?

* p(x, y) hoặc p(x, y|theta) và khi ta tiến hành predict cho y với giá trị x thì ta sẽ tính p(y|x) từ p(x, y)

## Gaussian Mixture Model dùng để làm gì?

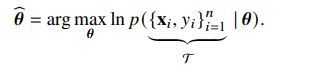
* **Khi train với data có label đầy đủ (supervised) -> linear / QDA**
* Với semi-supervised/unsupervised -> giải bài toán clustering (grouping similar x-values together in clusters)

## Giả thuyết cơ bản của GMM là gì?

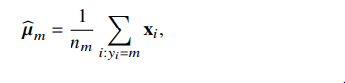
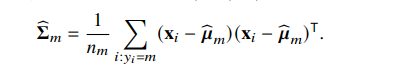
* Mỗi p(x|y) là một Gaussian distribution
* 

## Mô tả cách học 1 GMM

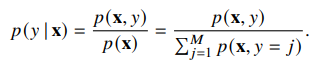
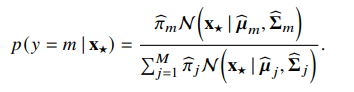
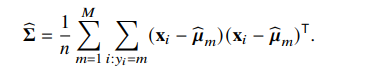
* Ta cần học tham số 
* Học GMM bằng cách maximize log-likelihood của bộ training data:



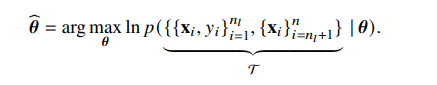
MLE ta thu được theta như sau:

* 
* 
* 

## Mô tả cách predict với GMM

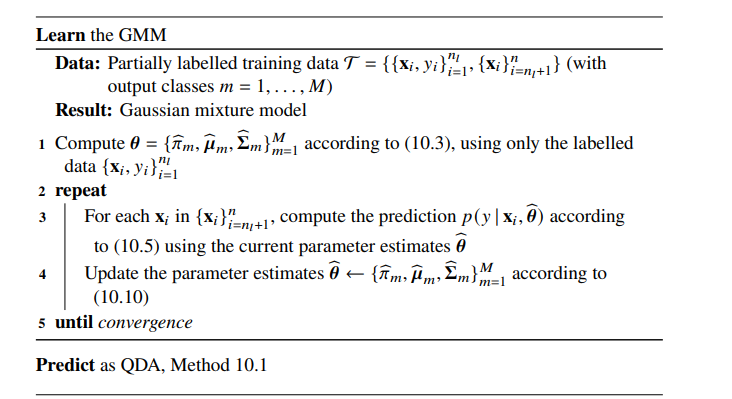
* Discriminant analysis:
  + Tính p(y|x) từ p(x, y):
  + 
  + Chọn hành động:
    - 
    - 
    - 
    - Do ln Gaussian là hàm bậc 2 của x => QDA
    - Giả sử tất cả covariance matrix băng nhau thì ta có chỉ duy nhất một covariance matrix để học => Ta được LDA
    - 

## Mô tả cách học semi-supervised learning với GMM

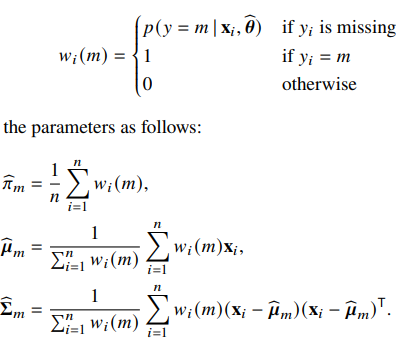
* MLE với theta sao cho:
* 

(không có closed form solution)

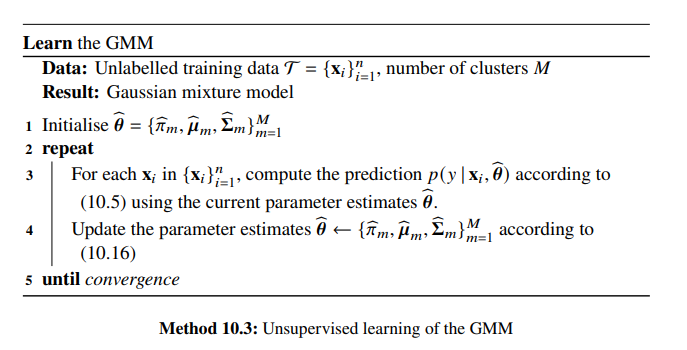
* Giải quyết:
  + Học GMM từ những data đã được label
  + Dùng GMM để predict missing output cho những data chưa được label
  + Cập nhật GMM với outputs được predict từ bước ii
  + Lặp lại bước ii, iii cho đến khi hội tụ



Update param với EM algorithm (Expectation-maximization):

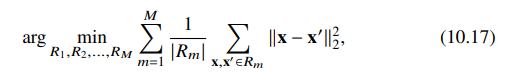


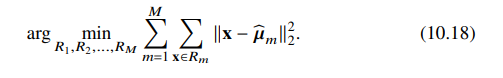
## Mô tả cách dùng GMM cho Cluster analysis



## Mô tả thuật toán k-means

* (i) Set cluster center to some values
* (ii) Determine which cluster Rm each xi belongs to, that is, find the cluster center that is closest to xi for all i = 1, …., n
* (iii) Update the cluster center as the average of all xi that belongs to Rm
* Lặp cho tới khi (ii) và (iii) hội tụ





* Chọn k - tuỳ vào người dùng nhưng thường dùng elbow method

# Chapter 11

## User aspects trong machine learning bao gồm những gì?

* Defining the ML problem: data (split), size of data, metrics, baseline
* Improving a ML model: dễ → khó, learning curve
* Nếu không thể collect được thêm data
* Những vấn đề thực tế với data
* Model có đáng tin hay không

## Làm sao để tạo một baseline để đo performance level?

Lấy cái đơn giản nhất có thể/human-level performance

## Làm thế nào để biết model có đáng tin hay không?

* Understand how certain prediction was made
* Worst case guarantee (tuy nhiên với model phức tạp thì sẽ khó -> đòi hỏi testing mô hình một cách cụ thể và kỹ lưỡng -> potential issues)

# Chapter 12

## Các vấn đề đạo đức trong ML bao gồm những gì?

* Sự công bằng của mô hình + hàm lỗi/tiêu chí so sánh:
* Với các tiêu chí (metrics) so sánh khác nhau, ta có thể đưa ra kết luận mô hình có công bằng hoặc không công bằng
* Không thể nào đạt được sự công bằng tuyệt đối với mọi metrics về mặt toán học

→ Fairness through awareness: khi report thì không được cherry-picking mà phải report toàn bộ, giúp cho mọi người hiểu được sự tổng quan

* Report kết quả không trung thực:
* Tạo ảo ảnh mô hình hoạt động rất tốt nhưng trên thực tế thì không phải như vậy
* Mô hình hoạt động chỉ số tốt, nhưng có thể con người thuần cũng đạt được gần tương tự
* Mô hình hoạt động chỉ số kém, nhưng lại đánh bóng kết quả tạo lầm tưởng mô hình đã đạt đến đỉnh cao
* Thiên hướng sử dụng các thuật toán có độ phức tạp cao hơn (RF over LR) + report với niềm tin rằng chúng sẽ hoạt động tốt hơn, nhưng thực tế thì méo phải như vậy
* Các kết quả không giải thích được (toàn các con số đơn thuần)
* Thiếu dữ liệu:
* Mô hình là một sự phản ánh của dữ liệu → nếu dữ liệu bị thiếu 1 nhóm thì mô hình sẽ hoạt động không tốt hoặc có những thiên kiến đối với các nhóm đó
* VD: một mô hình gen mặt người, da trắng > da đen thì khả năng cao là gen ra da trắng nhiều hơn.
* VD: gender bias in word2vec

## Ethics through awareness để chỉ điều gì?

Ethics through awareness: khi report thì không được cherry-picking mà phải report toàn bộ, giúp cho mọi người hiểu được độ hoạt động tổng quan của mô hình

Điều này là do với các tiêu chí (metrics) so sánh khác nhau, ta có thể đưa ra kết luận mô hình có công bằng hoặc không công bằng. Ngoài ra, không thể nào đạt được sự công bằng tuyệt đối với mọi metrics về mặt toán học.

Do đó, việc một mô hình có hoạt động một cách “đạo đức” hay không phụ thuộc hoàn toàn vào cách nhìn, cách đánh giá của của chúng ta.

## Hãy nêu 3 ví dụ các vấn đề đạo đức khi ứng dụng ML

* Với các góc nhìn khác nhau thì sẽ đưa ra các kết quả khác nhau:
* Swede/non-Swede example
* Báo cáo láo về sự hiệu quả/độ hoạt động của một mô hình:
* Elon Musk: xe tự lái cấp 5 trong khi mới được cấp 2
* Dữ liệu bị thiếu/dữ liệu giữa các nhóm có sự chênh lệch lớn:
* VD: mô hình sinh ảnh mặt người có nhiều dữ liệu huấn luyện là mặt thằng da trắng

## Có thể đạt được fairness sử dụng metrics nào?

Có rất nhiều metrics có thể sử dụng để đo sự fairness của một mô hình học máy, nhưng không có duy nhất một metrics nào có thể cho chúng ta sự fairness tổng quan.

Lí do là vì (ethics through awareness …)

Vì vậy, ta cần (ethics through awareness …)

## Có các kỹ thuật nào để phát hiện bias trong ML model?

* Kiểm tra dữ liệu: nếu dữ liệu bị skewed (có một nhóm hơn hẳn 1 nhóm khác, như nam > nữ, white > non-white) thì khả năng cao là mô hình sẽ bị bias

## Mô tả 1 hype trong ML và phân tích các khía cạnh đạo đức

Cái này dùng chatgpt nhé, nhiều lắm

## Tóm tắt

* **Fairness and error functions**

Fairness through awareness

* Tùy từng ứng dụng sẽ có các metrics về fairness khác nhau, nên khi report kết quả cần phải report toàn bộ, không chỉ chọn những metrics cụ thể

Complete fairness is mathematically impossible

* Không thể nào đạt tới sự “fair” tuyệt đối, trừ khi data hoàn toàn sạch
* **Misleading claims about performance**

Criminal sentencing

* Mô hình không thực sự hoạt động tốt hơn con người là bao

Explaining model in an understandable way

* Tùy vào ngữ cảnh mà chúng ta cần giải thích số liệu sao cho hiệu quả

Case study: Cambridge Analytica

* Report láo kết quả, sử dụng AI hype để truyền thông bẩn cho thuật toán kém hiệu quả của họ

Case study: Medical Imaging

* Có thiên hướng sử dụng các phương pháp phức tạp vì niềm tin rằng chúng đem lại “sự hiệu quả” tốt hơn, nhưng thực tế thì không phải như vậy
* **Limitations of training data**

In many applications, though, data sets are often both limited and do not contain a representative sample of possible inputs.

+ Skewed data: more images with white people, more texts with “socially biased” contents