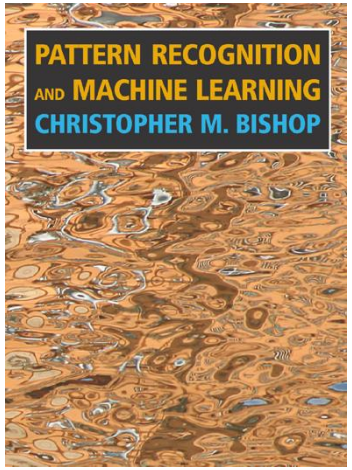
Linear Regression in Pharmacology

Mohammad Ghazi Vakili, PhD

Pharmacology Department

Applying Statistical Learning to Drug Discovery and Clinical Data

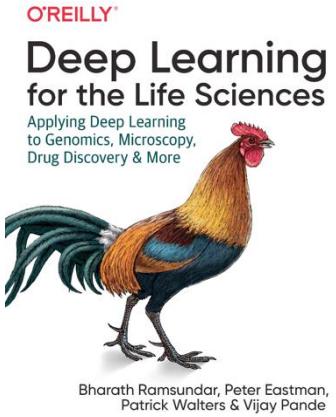
Machine learning algorithms



Pattern Recognition and Machine Learning

by Christopher M. Bishop

✓Chapter 3: Linear Model for Regression



Deep Learning for the Life Sciences

by Bharath Ramsundar, Peter Eastman, and Vijay S. Pande

✓Chapter 2: Introduction to Deep Learning

Prerequisite: Calculus I

Course structure

- **Weekly Lectures:**
  Introduce core ML concepts in Pharmacology applications
- **Python Lab:**
  Weekly Hands-on coding to implement algorithms and visualize data
- **Quizzes (30%):**
  Focus on understanding *when and why* to use each ML algorithm
- **Project (40%): :**
  Build an end-to-end ML pipeline for a real Pharmacology dataset
- **Exam (30%): :**
  Assess both conceptual understanding and practical application

Machine learning algorithms

What we will cover today's lecture

- Why Machine Leaning
- QSAR
- QSAR and Linear Regression
- Example and Exercise

Why LR in Pharmacology

- Pharmacology is data-rich
- Need to predict potency, toxicity, and Pharmacokinetics.
- Machine Learning (ML) enables:
  - Discovery of hidden patterns
  - Quantitative prediction of biological activity
  - Linking chemical structure → pharmacological behavior
- Next is Deep Learning

What is QSAR

QSAR (Quantitative Structure–Activity Relationship):
A mathematical model that relates molecular **structure** to biological **activity**

- **Hansch & Fujita (1964):**
- Led to famous **multivariate regression models** in pharmacology
- First linear equation:

$$\log\left(\frac{1}{c}\right) = a * logP + b * \sigma + c * MW + d$$

What is QSAR

QSAR in Practice: A Simple Example

Predicting Inhibitor Potency (pIC_{50}) for a Protease Target

Compound	Mol. Wt. (Da)	LogP	pIC_{50} (activity)
A	250	2.1	6.2
B	280	2.8	6.9
C	310	3.2	7.1
D	340	3.9	7.5
E	320	3.2	??

Question: Can we build a model to predict pIC_{50} for new compounds?

Machine learning algorithms

Supervised learning

- Data include **input features** and known **output labels**
  - Model learns a mapping: $f(x) \rightarrow y$
- Goal: **predict or estimate** outcomes for new data
- Examples in pharmacology:
  - Predicting IC_{50} (regression)
  - Classifying compounds as toxic/non-toxic (classification)

Algorithms:
Linear Regression, Logistic Regression, Decision Trees, Neural Networks

Unsupervised learning

- Data have **no predefined labels**
  - Model finds hidden structure or patterns
- Goal: **group** or **reduce** data complexity
- Examples in pharmacology:
  - Clustering compounds by chemical similarity
  - Grouping patients by response profile
  - Dimensionality reduction of molecular descriptors (e.g., PCA)

Algorithms:
k-Means, Hierarchical Clustering, PCA, Autoencoders

Mathematics Recap

Vector and Matrix

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

Matrix

$$xA = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} x_1 a_{11} + x_2 a_{21} & x_1 a_{12} + x_2 a_{22} \end{bmatrix}$$

$$B = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad (B)^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

$$B^T = \begin{bmatrix} a & c \\ b & d \end{bmatrix}$$

Linear Regression

Chapter 3

Linear Regression:

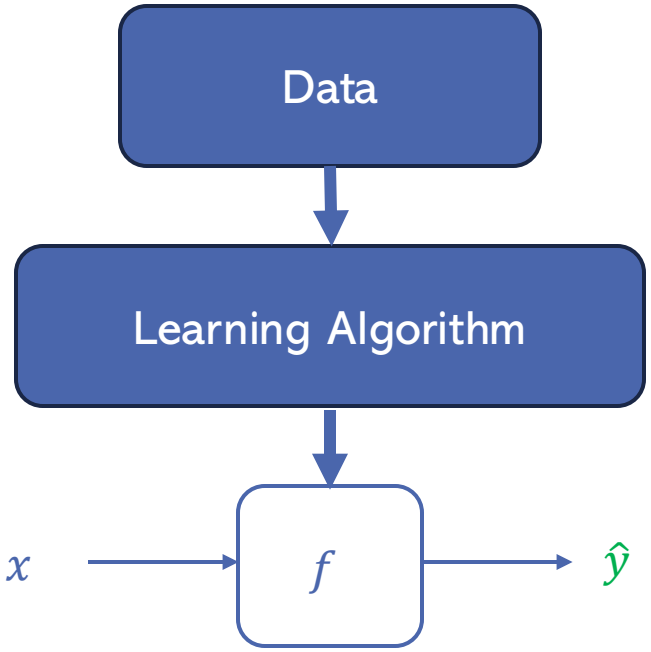
Training set: Data used to train the model

Compound	Mol. Wt. (Da)	LogP	pIC_{50} (activity)
A	250	2.1	6.2
B	280	2.8	6.9
C	310	3.2	7.1
D	340	3.9	7.5
…	…	…	…
E	320	3.2	??

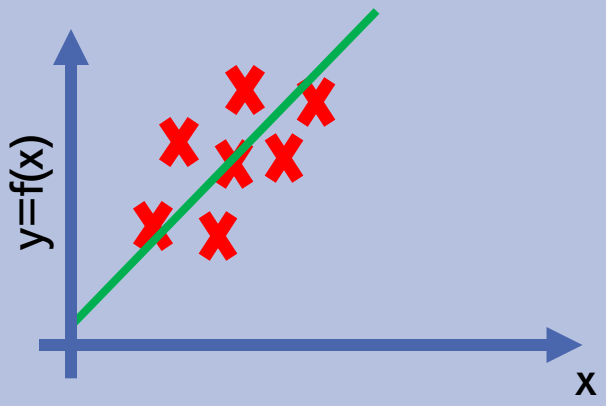
x : Input (features)

y: output (target)

Linear Regression



How to represent f



One size variable

$$f_{w,b}(x) = wx + b$$

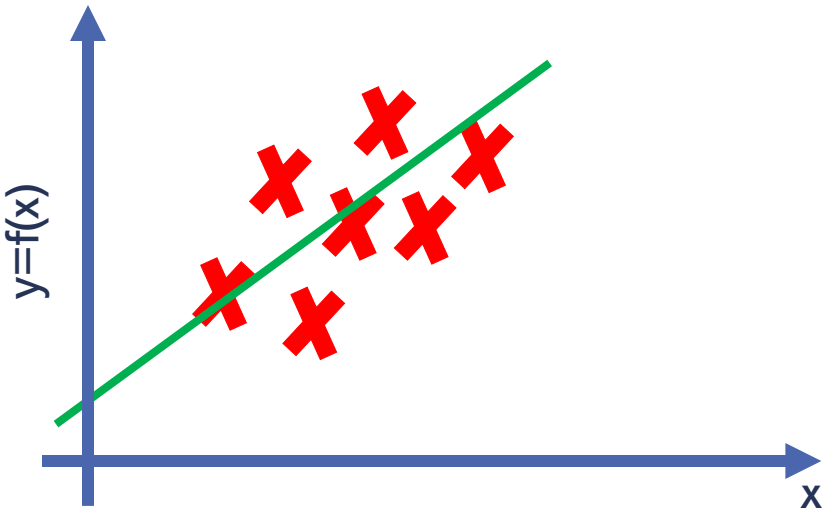
Multi variable

$$f_{w,b}(x) = w_0 + w_1x_1 + w_2x_2 + w_3x_3 =$$

$$w_0 + \sum_{i=0}^{3} w_i x_i$$

Linear Regression: Model

Cost Function $J(w,b)$ { Measures the average squared difference between predicted $(\hat{y})$ and actual (y) values. The goal is to find parameters w, b that **minimize this error**.



$\hat{y} = f_{w,b}(x^{(i)})$

$f_{w,b}(x) = w_0 + wx^{(i)}$

$$J(w) = \frac{1}{2m}\sum_{i=0}^{m}(\hat{y}^{(i)} - y^{(i)})^2$$

With :

$$\hat{y} = f_{w,b}(x^{(i)}) = w_0 + wx^{(i)}$$

Find the parameters w and b that **minimize the cost function $J(w,b)$**

Closed Form (Normal Equation)

To minimize the $J(w)$ take the derivative with respect to w and set it to zero:

$$\frac{\partial J}{\partial w} = \left(\frac{1}{m}\right)\Phi^{\mathrm{T}}(\Phi w - y) = 0$$

Where:

$$\hat{y} = wx + b$$

$$\hat{y} = w_1 x + w_0 = [w_0\, w_1]\begin{bmatrix}1\\ x\end{bmatrix} = w\phi$$

Simplify

$$\phi^T \phi w = \phi^T y$$

Now solve w

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

This is called the **Normal Equation,** it gives the exact, closed-form solution.

Note:

- Requires invertible $\phi^T \phi w$ (issue with multicollinearity).
- Gradient Decent will be use for higher dimension (Training)

The Coefficient of Determination

Measures how well our regression line fits the data

$$R^2 = 1 - \left(\frac{SS_{res}}{SS_{tot}}\right)$$

$SS_{res} = \sum(y_i - \hat{y_i})^2$ → unexplained variation (residuals)

$SS_{tot} = \sum(y_i - \overline{y_i})^2$ → total variation in data

Interpretation:

- $R^2 = 1 \rightarrow$ Perfect fit (model explains all variation)
- $R^2 = 0 \rightarrow$ Model explains none of the variation
- Example: $R^2 = 0.85$ means the model explains **85% of the change** in the response.

Recap

- QSAR

$$\hat{Y} = \log\left(\frac{1}{c}\right) = a * logP + b * \sigma + c * MW + d$$

- Linear Regression

$$\hat{y} = w_1 x + w_0 = [w_0\ w_1]\begin{bmatrix}1\\ x\end{bmatrix} = w\phi$$

$$J(w) = \frac{1}{2m}\sum_{i=0}^{m}(\hat{y}^{(i)} - y^{(i)})^2$$

- How to compute the coefficients
  - Normal Equation

$$w = (\phi^T\phi)^{-1}\phi^T y \quad \text{and} \quad R^2 = 1 - \left(\frac{SS_{res}}{SS_{tot}}\right)$$

Normal Function Example:

Fit Linear model to

No.	x	y
1	1	2
2	2	3
3	3	5

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \qquad y = \begin{bmatrix} 2 \\ 3 \\ 5 \end{bmatrix}$$

$$w = (X^T X)^{-1} X^T y$$

$$X^T X = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix} \qquad X^T y = \begin{bmatrix} 10 \\ 23 \end{bmatrix}$$

$$(X^T X)^{-1} = \frac{1}{6} \begin{bmatrix} 14 & -6 \\ -6 & 3 \end{bmatrix}$$

$$w = \begin{bmatrix} 0.33 \\ 1.5 \end{bmatrix} \qquad y = 0.33 + 1.5x$$

$$R^2 = 0.964$$

Lab Exercise 1

Predicting Drug-Induced Heart Rate Response Using Multi-Feature Regression

	Dose (mg/kg)	Weight (kg)	Age (years)	Response (Δ bpm)
1	3.62	74.3	42.49	9.05
2	7.66	56.82	37.21	21.94
3	6.12	52.6	62.29	17.46
4	5.19	87.96	41.05	16.55
5	2.09	88.63	37.64	2.71
6	2.09	82.34	49.42	5.51
7	1.41	62.18	31.34	-3.75
8	7.06	53.91	61.1	21.69
9	5.21	77.37	28.35	14.66

Develop a multiple linear regression model that predicts the heart rate response (Δ bpm) based on the following independent variables:

- Dose (mg/kg): Drug dosage per kilogram of body weight
- Weight (kg): Subject's total body mass
- Age (years): Subject's age

The goal is to **quantify the contribution of each factor** to the observed pharmacological response and interpret the coefficients in a biological context.

Lab Exercise 1

Predicting Drug-Induced Heart Rate Response Using Multi-Feature Regression

	Dose (mg/kg)	Weight (kg)	Age (years)	Response (Δ bpm)
1	3.62	74.3	42.49	9.05
2	7.66	56.82	37.21	21.94
3	6.12	52.6	62.29	17.46
4	5.19	87.96	41.05	16.55
5	2.09	88.63	37.64	2.71
6	2.09	82.34	49.42	5.51
7	1.41	62.18	31.34	-3.75
8	7.06	53.91	61.1	21.69
9	5.21	77.37	28.35	14.66

$$\hat{Y} = \mathrm{w}_0 + w_1 * \mathrm{Dose} + w_2 * Weight + w_3 * Age$$

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

Solution:

$$\hat{Y} = -3.231 + 3.410 * \mathrm{Dose} + 0.003 * Weight - 0.023 * Age \text{ and } R^2 = 0.99$$

Link to Google Colab

Next session

- Gradient Decent
- Python
- Python implementation of the LR
  - Normal Equation
- Non-Linear LR
  - Gaussian (Radial Basis) Regression
  - Logarithmic and Exponential Models
  - Generalized Linear Models (GLMs)
  - Sigmoidal / Logistic Regression (Hill Equation)

Appendix A

QSAR Parameters

$$\log\left(\frac{1}{c}\right) = a * logP + b * \sigma + c * MW + d$$

Symbol	Description	Pharmacological Interpretation
C	Molar concentration required to produce a defined biological effect	Inversely related to potency (lower C → higher activity)
log(1/C)	Activity term (higher = more active)	Used as dependent variable (Y)
log P	Partition coefficient (lipophilicity between octanol and water)	Measures hydrophobicity; affects membrane permeability and receptor binding
σ (sigma)	Hammett substituent constant	Represents electronic effects of substituents on the aromatic ring (electron-withdrawing or -donating)
MW	Molecular weight	Captures steric or size-related contribution to activity
a, b, c	Regression coefficients	Quantify how each property influences activity (sign and magnitude indicate direction and strength)
d	Intercept	Baseline biological activity independent of descriptors

Linear Regression: Model

We want to find the best-fitting line that predicts $\hat{y}$ from x:

$$\hat{y} = wx + b$$
$$\hat{y} = w_1 x + w_0 = [w_0\ w_1] \begin{bmatrix} 1 \\ x \end{bmatrix}$$

or, in vector/matrix form (for many samples and features):

$$\hat{y} = w\phi$$

Where

- ϕ= design matrix (rows = samples, columns = features, plus 1 for intercept),
- w = weights (parameters),
- $\hat{y}$ = predicted values.

Gradient Descent (When D or N is large)

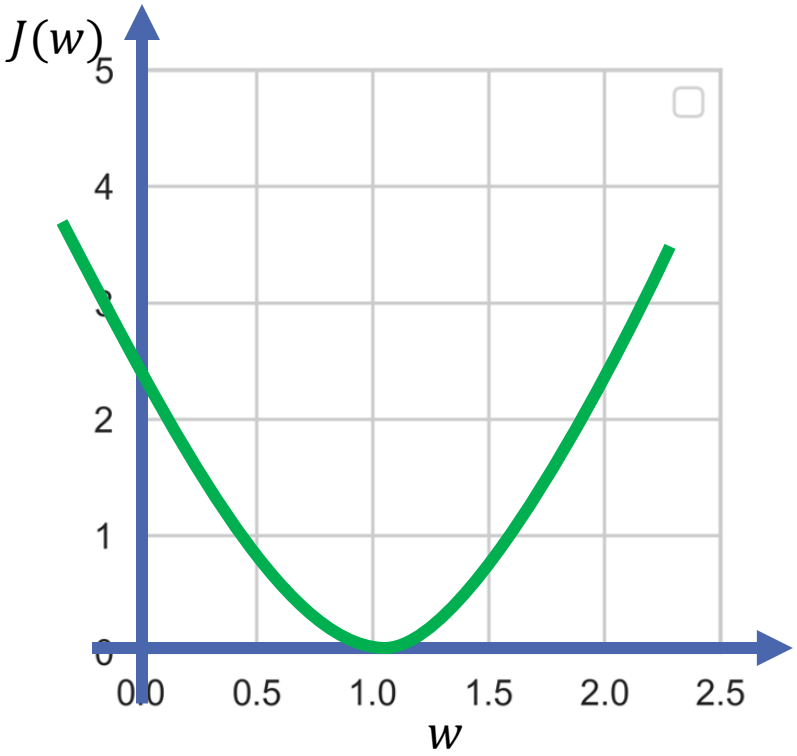
This **iterative optimization** steps is called Gradient Decent Instead of computing the $(\phi^T\phi)^{-1}\phi^T y$, it **updates weights step-by-step** by following the gradient (slope) of the cost function:

$$w := w - \alpha \frac{\partial J}{\partial w}$$

where α is the **learning rate.**

$$\frac{\partial J}{\partial w} = \left(\frac{1}{m}\right) \Sigma \left(\hat{y}^{(i)} - y^{(i)}\right) x^{(i)}$$

- Scales to streaming data; no matrix inversion.
- Learning rate α controls convergence speed/stability.
- Batch / stochastic / mini-batch variants.



Gradient Descent (When D or N is large)