Statistics and Machine Learning

Part 1: Theory and Regression Problems

Conrado Miranda
LBiC, FEEC, University of Campinas

Contents

Introduction

Why statistics?

- Everything needs a **model** and statistics provide a framework to deal with **data**.
- Most **machine learning** algorithms are based on **statistics** and **graph theory**.

Focus of the presentation

- Understanding statistics properties;
- Building and expanding models;
- Solving regression, classification, and clustering problems.

Random Variable: Definition

Definition

A random variable (rv) $X : \Omega \to E$ is a measurable function from the set of possible outcomes Ω to some set E.

Example: coin toss bet

$$X(\omega) = \begin{cases} 1, & \omega = H \\ -1, & \omega = T \end{cases} \tag{1}$$

Random Variable: Definition

Function of random variable

Let $f(X)\colon E_X \to E_Y$, then a new random variable Y may be defined as:

$$Y = f(X) = f \circ X, \quad Y : \Omega \to E_y \tag{2}$$

Realization

A realization of a rv X is the value x that is actually observed when the variable is measured, and it's denoted as:

$$x \sim X \tag{3}$$

Random Variable: Discrete and Continuous Distributions

Discrete

A discrete probability distribution $D\colon \sigma(\Omega) \to [0,1]$ is described by its probability mass function (pmf) $p(X=x)$, such that:

$$\sum_{x\in\Omega} p(X=x) = 1, \quad p(X=x) \geq 0 \tag{4}$$

Continuous

A continuous probability distribution $D\colon \sigma(\Omega) \to [0,1]$ is described by its probability density function (pdf) $p_X(x)$, such that:

$$\int_\Omega p_X(x)\mathrm{d}x = 1, \quad p_X(x) \geq 0 \tag{5}$$

Random Variable: Expectation

Definition

The expectation operator is the average of values a function achieves for each event, pondered by the probability of the event:

$$\mathbb{E}_{x\sim X}[f(x)] = \sum_{x\in\Omega} f(x)p(X = x) \tag{6a}$$

$$\mathbb{E}_{x\sim X}[f(x)] = \int_{\Omega} f(x)p_X(x)\mathrm{d}x \tag{6b}$$

Random Variable: Expectation (applications)

Mean

$$\mu_X = \mathbb{E}_{x \sim X}[x] = \sum_{x \in \Omega} x \; p(X = x) \tag{7}$$

Variance

$$\sigma_X^2 = \mathbb{E}_{x \sim X}[(x - \mu_X)^2] \tag{8}$$

Entropy

$$H_X = \mathbb{E}_{x \sim X}[-\log x] \tag{9}$$

Random Variable: Example

Problem

Let p the probability of a coin toss providing heads H, and $1 - p$ the probability of tails T. Let X be a bet that pays 1 if the coin lands on H, and charges -1 if it lands on T. Determine the expected pay-off.

Random variables definition

Coin toss distribution: $p(H) = p$, $p(T) = 1 - p$.
Pay function: $f(H) = 1$, $f(T) = -1$.

Expected pay-off

$$V = \mathbb{E}_{c \sim C}[f(c)] = \sum_{c \in \{H,T\}} f(c)p(c) = p - (1-p) = 2p - 1 \quad (10)$$

Main Distributions: Categorical

If a realization of the random variable X has to one of k values, then its distribution D is the categorical distribution. If p_i is the probability of obtaining the i-th value, then

$$p(X = x_i; \{p_i\}) = p(\mathbf{X}; \{p_i\}) = \prod_{i=1}^{k} p_i^{X_i}, \quad \sum_{i=1}^{k} p_i = 1, p_i \geq 0 \quad (11)$$

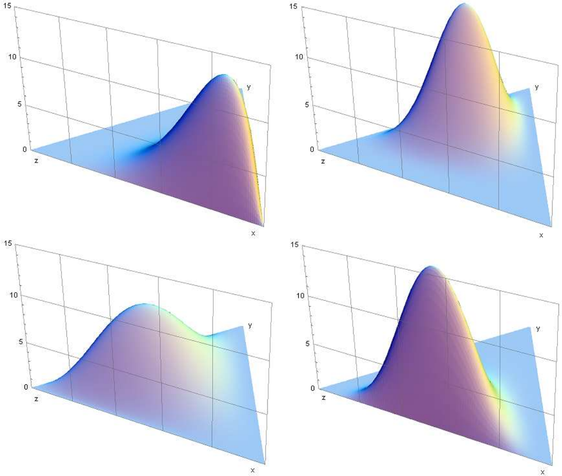
where $\mathbf{X} = [X_1, X_2, \ldots, X_k]$ and $X_i = 1[X = x_i]$.

Examples

Coin flip, dice roll, roulette, card games.

Main Distributions: Dirichlet

$$S^{d-1} = \{\theta \in \mathbb{R}^d | \theta_i \geq 0, \sum_i \theta_i = 1\} \tag{12}$$



Main Distributions: Normal and Laplace

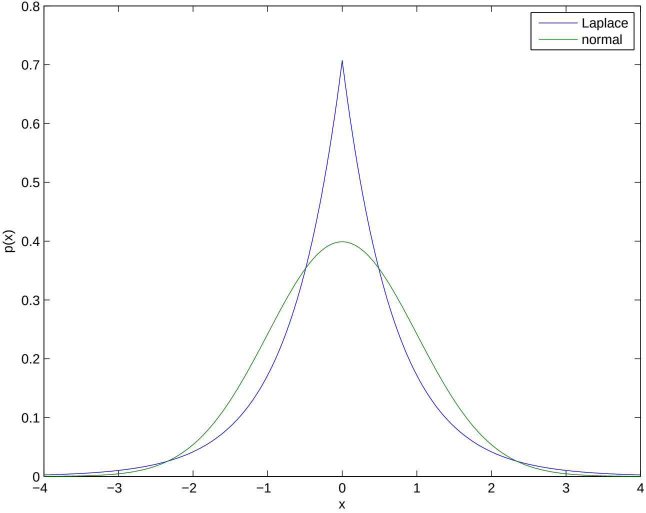
Normal distribution

$$p(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{13}$$

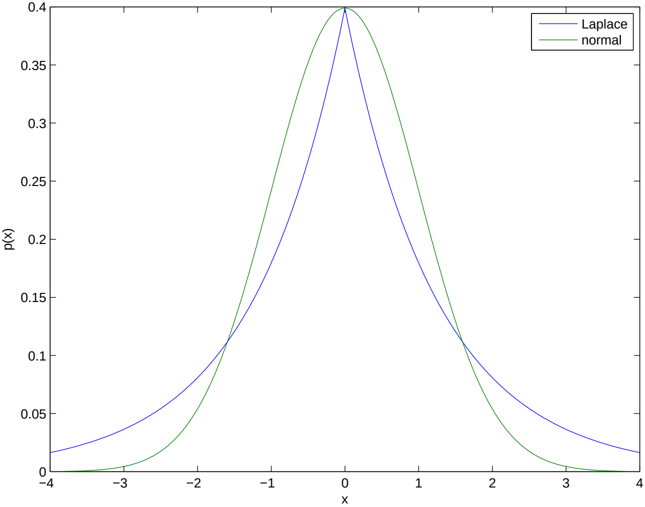
Laplace distribution

$$p(x; \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right) \tag{14}$$

Main Distributions: same variance



Main Distributions: same maximum



Distribution Properties: Multiple Variables

- Joint and marginal distributions

$$\underbrace{p(x_1, x_2, y_1, y_2)}_{\text{joint distribution}} = \underbrace{p(x_1, x_2 | y_1, y_2)}_{\text{marginal distribution}} \; p(y_1|y_2)p(y_2) \tag{15}$$

- Marginalization

$$p(x) = \sum_y p(x, y) \tag{16}$$

- If X is independent of Z given Y, then

$$p(x|y, z) = p(x|y) \tag{17}$$

- Bayes' Theorem

$$p(x|y, z) = \frac{p(y|x, z)p(x|z)}{p(y|z)} \quad p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \tag{18}$$

Distribution Properties: Sufficient Statistics

Statistic

A statistic is the application of a function to a sample set.
Example: sample mean.

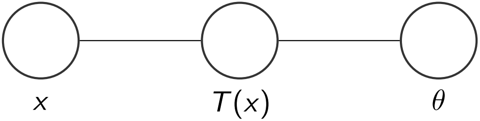
$$\overline{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{19}$$

Definition

A statistics $T(x)$ is sufficient for the underlying parameter θ if

$$p(X = x|T(x) = t, \theta) = p(X = x|T(x) = t) \Leftrightarrow p(\theta|t, x) = p(\theta|t) \tag{20}$$

Distribution Properties: Factorization



Fisher-Neyman factorization theorem

$T(x)$ is a sufficient statistics iff

$$p(x|\theta) = h(x)g(\theta, T(x)) \tag{21}$$

Exponential family

$$p(x|\theta) = h(x)\exp(\eta(\theta) \cdot T(x) - A(\theta)) \tag{22}$$

Distribution Properties: Central Limit Theorem

Lindeberg-Lévy Central Limit Theorem

Let X_i, $i = \{1, \dots, n\}$, be independent and identically distributed (iid) random variables with $\mathbb{E}[X_i] = \mu$ and $Var[X_i] = \sigma^2 < \infty$. Then, as $n \to \infty$,

$$\sqrt{n}\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) - \mu\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \tag{23}$$

Estimators: Maximum Likelihood Estimator (MLE)

Likelihood

Let $\{x_i\}$ be iid samples from a distribution with parameter θ. The likelihood of θ is defined by:

$$\mathcal{L}(\theta|x_1, \ldots, x_n) = p(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta) \qquad (24)$$

Estimator

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \sum_{i=1}^{n} \log p(x_i|\theta) \qquad (25)$$

Estimators: Maximum A Posteriori (MAP)

A Posteriori Probability

$$p(\theta|x_1, \ldots, x_n) = \frac{p(\theta)\prod_{i=1}^{n} p(x_i|\theta)}{p(x_1, \ldots, x_n)} \tag{26}$$

Estimator

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \log p(\theta) + \sum_{i=1}^{n} \log p(x_i|\theta) \tag{27}$$

Relationship to MLE

$$p(\theta) = C \Rightarrow \hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{MLE}} \tag{28}$$

Estimators: KL-Divergence

KL-Divergence

$$KL[p(x)||q(x)] = \int p(x) \log \frac{p(x)}{q(x)} dx \quad (29a)$$

$$= -H_{p(x)} - \mathbb{E}_{x \sim p(x)}[\log q(x)] \quad (29b)$$

Empirical distribution

$$p_s(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i) \quad (30a)$$

$$\mathbb{E}_{x \sim p_s(x)}[\log p(X, \theta)] = \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i) \log p(x, \theta) \quad (30b)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \log p(x_i, \theta) \quad (30c)$$

Loss Functions

Objective

$$\min_y \mathbb{E}_{x \sim X}[L(x, y)] \tag{31}$$

Common losses

- $L_{sq}(x, y) = (x - y)^2$ means predicting the mean of x.
- $L_{av}(x, y) = |x - y|$ means predicting the median of x.
- $L_{\log}(x, y) = \log(1/p(z = x)), z \sim y$, means minimizing the description length of x.
- $L_{bi}(x, y) = -y^x(1 - y)^{(1-x)}$ means predicting the probability of x happening.

Regression: Standard Linear

Mathematical description

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|_2^2 \tag{32}$$

Statistical description*

$$\hat{\beta} = \hat{\beta}_{MLE} = \arg\max_{\beta} \sum_{i=1}^{n} \log p_{\mathcal{N}}(y_i; x_i\beta, I) \tag{33}$$

Regression: Ridge Linear

Mathematical description

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2 \quad (34)$$

Statistical description*

$$\hat{\beta} = \hat{\beta}_{MAP} = \arg\max_{\beta} \log p_{\mathcal{N}}(\beta; 0, \lambda^{-\frac{1}{2}} I) + \sum_{i=1}^{n} \log p_{\mathcal{N}}(y_i; x_i\beta, I) \quad (35)$$

Regression: Lasso Linear

Mathematical description

$$\hat{\beta} = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{36}$$

Statistical description*

$$\hat{\beta} = \hat{\beta}_{MAP} = \arg\max_{\beta} \log p_{\mathcal{L}}(\beta; 0, \lambda^{-1} I) + \sum_{i=1}^{n} \log p_{\mathcal{N}}(y_i; x_i\beta, I) \tag{37}$$

Regression: Other Regressors

Generalization

Just replace $x_i\beta$ with $f(x_i, \beta)$, where $f(\cdot)$ is a general learner.

Example: matrix decomposition

Find $V \in \mathbb{R}^{n\times k}$ and $U \in \mathbb{R}^{k\times m}$ to fit $X \in \mathbb{R}^{n\times m}$.

1. To approximate full X, a normal error is considered.
2. To avoid overfitting, assume each value of V and U comes from a normal distribution.

$$X_{i,j} - V_{i,:}U_{:,j} = E_{i,j} \sim \mathcal{N}(0,1), V_{i,j} \sim \mathcal{N}(0,1), U_{i,j} \sim \mathcal{N}(0,1) \tag{38a}$$

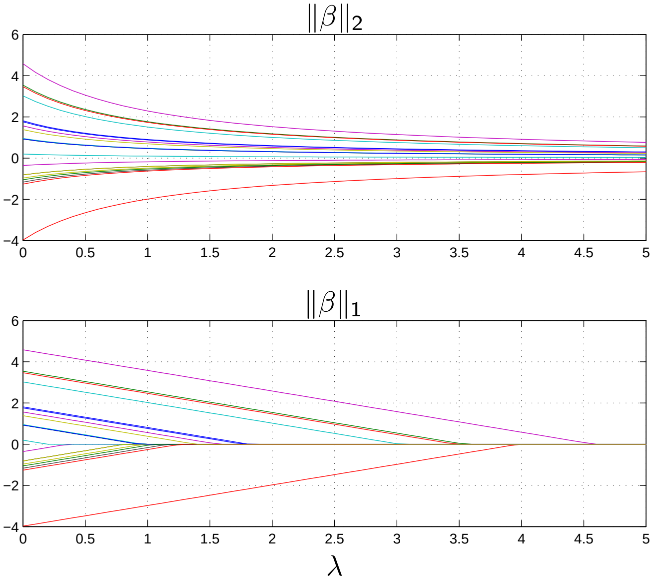
$$\min_{V,U} \|X - VU\|_2^2 + \|V\|_2^2 + \|U\|_2^2 \tag{38b}$$

Example: regularized linear regression

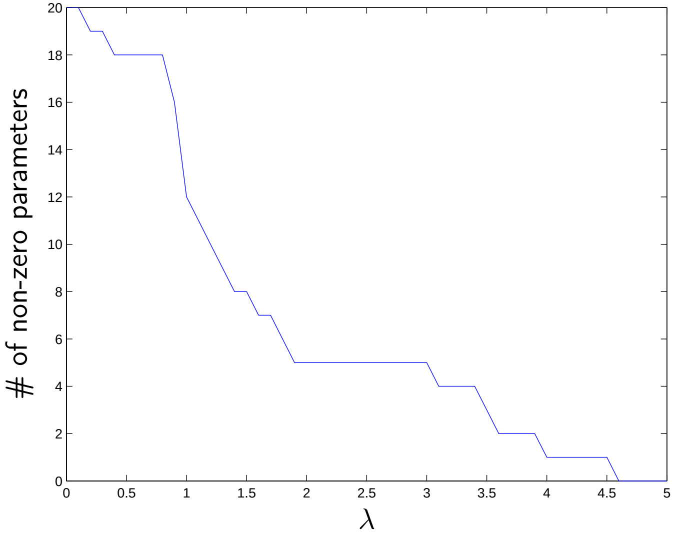
Parameters

- # of variables: 20
- # of training samples: 100
- # of test samples: 100k
- Noise: $\mathcal{N}(0, 1)$
- Features orthonormalized for easier solution

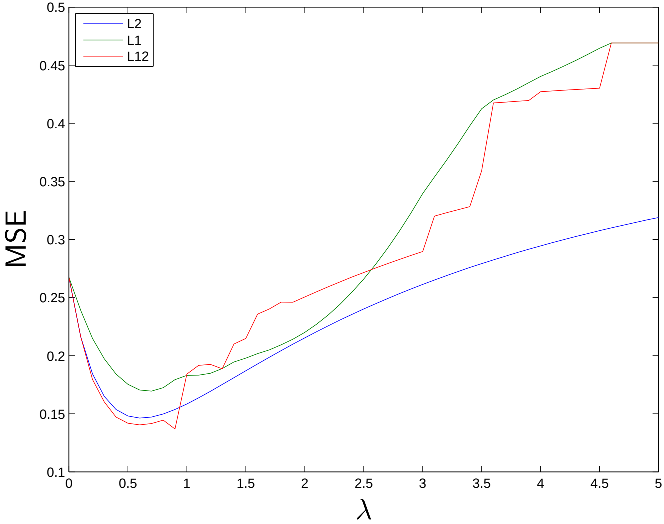
Example: parameters



Example: number of parameters



Example: noiseless test error



Conclusion

Statistics is very useful, but can be hard

- Can be used to propose new models;
- Lots of properties and relationships;
- Responsible for some (but not all!) important machine learning improvements.

What to expect from part 2

- How to describe models visually to ease understanding;
- Examples of algorithms for problems of classification and clustering.