Lecture1 probability

Recall from last time

- Experiment: is an activity or procedure that produces distinct, well-defined possibilities called **outcomes**.
- The set of all outcomes is called the **sample space**, and is denoted by Ω.
- Trial: doing the experiment once and getting an outcome.
- The subsets of Ω are called **events** events.
- Given an outcome $\omega \in \Omega$ we say that the event $E \subset \Omega$ **occured** if $\omega \in E$.

Lecture2 Random Variables

Recall from last time

- A random variable is a function from the sample space to a number (or vector).
- A distribution function is $F(x) = \mathbb{P}(X \leq x)$.
- A discrete random variable takes discrete values, i.e. $0, 1, 2, 3, \ldots$. The probability mass function is defined as $f(x) = \mathbb{P}(X = x)$.
- A random variable is called continuous if the distribution function F can be written as

$$F(x) = \int_{-\infty}^{x} f(v)dv$$

  for a piecewise continuous function f. f is called the density function.

Lecture3 Random Variables

Recall from last time

- The Joint Distribution Function for $X = (X_1, \dots, X_n)$ is the function

$$F(x) = \mathbb{P}(X_1 \leq x_1; \dots; X_n \leq x_n) \quad x = (x_1, \dots, x_n)$$

- Random variables $Z = (X, Y)$ are said to be independent if

$$F(z) = F(X \leq x)F(Y \leq y) \quad z = (x, y).$$

- A sequence of random variables $X_1, \dots, X_n$ is simply a random vector $X = (X_1, \dots, X_n)$
- The sequence is independent if $F_{X_1,\dots,X_n}(x_1, \dots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$.
- The sequence is identically distributed if $F_{X_i} = F_{X_j}$.
- If both then IID (Independent and Identically Distributed)

Lecture4 Concentration

Recall from last time

- Concentration of measure is a statement of the form, for every $0 < \delta < 1$ there is an $\epsilon > 0$ such that

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon) \leq \delta$$

- Chebyschev, we only know variance

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq \epsilon) \leq \frac{\mathbb{V}(X)}{n\epsilon^2}$$

- Hoeffding, we only know boundedness $a \leq X \leq b$

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq \epsilon) \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}}.$$

Recall from last time

- Bennett, we know boundedness and variance

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[\overline{X}_n]| \geq \epsilon) \leq 2\exp\left(-\frac{n\sigma^2}{b^2} h\left(\frac{b\epsilon}{\sigma^2}\right)\right)$$

  where $h(u) = (1+u)\log(1+u) - u$ for $u > 0$.
- A confidence interval is a random interval I that is determined from $X_1, \ldots, X_n$ and satisfies

$$\mathbb{P}(\mathbb{E}[X] \in I) \geq 1 - \delta$$

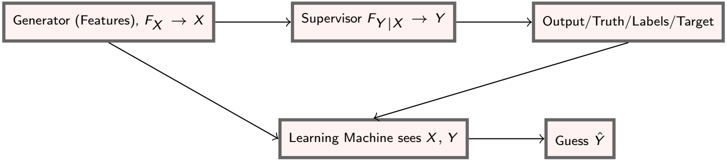
- The confidence $1 - \delta$ tells us that **before** we compute the interval, the probability that our interval I covers $\mathbb{E}[X]$ is at least $1 - \delta$.

Lecture5 Risk

Recall from last time

Setup

1. The generator of the data G
2. The supervisor S
3. The learning machine LM.



Recall from last time

- Statistical model (Our assumptions of the truth)
- The model space $\mathcal{M}$, what the learning machine searches in.
- The loss function L measuring the performance of a function $g \in \mathcal{M}$ w.r.t data.
- The risk which is expected loss.
- The main objective of the learning machine is to find $\hat{g} \in \mathcal{M}$ that minimizes risk.

Recall from last time

- We talked about the following learning problems
  - Find f
  - Regression
  - Pattern recognition
- We defined the regression function

$$r(X) = \mathbb{E}[Y \mid X]$$

which is the target to hit with Regression.

Lecture6 Estimation

Recall from last time

- The Data is our random variables $X = (X_1, \ldots, X_n)$. The data is an observation of our Data.
- A statistic is a function $\hat{\theta}$ from the data-space. We apply it on X, namely we are interested in $\hat{\theta}(X)$.
- An estimator is a statistic that is supposed to "estimate" an unknown quantity, say θ^*. Therefore we can speak about bias

$$\text{bias} = \mathbb{E}[\hat{\theta}(X)] - \theta^*$$

- A simple measure of performance is the standard deviation of the estimator, called, the standard error.
- The risk of the estimator w.r.t the quadratic loss can be decomposed as

$$\mathbb{E}[(\hat{\theta}(X) - \theta^*)^2] = (\text{bias}(\hat{\theta}))^2 + (\text{se}(\hat{\theta}))^2$$

Recall from last time

- We also defined different modes of convergence
  - Almost sure convergence
  - Convergence in probability
  - Convergence in distribution (we did not define this)
- an estimator is asymptotically consistent if it converges in probability to the true value.

Lecture7 Estimation Risk

Recall from last time

- We saw an example of different ways to construct estimators for a problem, and we calculated their standard errors. All estimators are not created equal.
- We explored the log-Loss, i.e. $L(z, \alpha) = -\ln p_\alpha(z)$, where p_α is a proposal density for our data, we assume that there is an α^* such that the data comes from p_{α^*}.
- We saw that the empirical risk is the negative log Likelihood

$$\hat{R}(\alpha) := \frac{1}{n}\sum_{i=1}^{n}(-\ln(p_\alpha(X_i)))$$
$$R(\alpha) = \mathbb{E}[-\ln(p_\alpha(X))]$$

Recall from last time

- We explored the problem of estimating the σ in $N(0, \sigma^2)$ using the Likelihood.
- We considered the conditional likelihood, i.e. our proposal density is of the form $f_\alpha(x, y) = p_\alpha(y \mid x)p(x)$ for some fixed $p(x)$.
- We saw
  - $p_{\alpha^*, X} = N(\alpha_1 X + \alpha_2, \alpha_3^2)$, Linear regression
  - $p_{\alpha^*, X} = \text{Bernoulli}(G(\alpha_1 X + \alpha_2))$,

    $$G(x) = \frac{1}{1 + e^{-x}}$$

    Logistic regression

Lecture8 Generating Random Variables Markov Chains

Recall from last time

- We explored how a computer which is fully deterministic, can produce something that looks random, i.e. pseudorandom.
- Pseudorandom sequences
- Period of a dynamical system
- We explored ways to go from our rudimentary dynamical sequence to something which is uniform $[0, 1]$.
- Linear Congruential Generators
- Bernoulli, discrete, shuffling
- Permutation test

Lecture9 Markov Chains

Recall from last time

- We defined a stochastic process as a index family of random variables X_α where $\alpha \in I$ is the index set. The base example is a sequence of i.i.d. random variables, $X_1, \ldots, X_n$, here the index set is $\mathbb{N}$.
- We defined the concept of Markov chain, which is a stochastic process which takes a finite number of states and its dependency on the past is only the previous value, i.e.

$$\mathbb{P}(X_{t+1} = x \mid X_1, \ldots, X_t) = \mathbb{P}(X_{t+1} = x \mid X_t).$$

Recall from last time

- We defined a homogeneous Markov chain, as one where the transition probabilities does not depend on the index t (or time).

$$\mathbb{P}(X_{t+1} = y \mid X_t = x) = P_{xy}$$

  that is the probability of transitioning from state x to state y does not depend on the particular time t.
- Since $X_t \in \mathbb{X}$, where $\mathbb{X}$ is called the state space and $\mathbb{X}$ is a finite set. We can define a matrix P_{xy} where $x, y \in \mathbb{X}$. We call this matrix the transition matrix.
- We can use the transition matrix to compute how distributions change when the Markov chain progresses. I.e. let us assume that at time t we have a distribution over the states p_t, then the distribution at time $t+1$ is computed as

$$p_{t+1} = p_t P \qquad p_t = p_0 P^t.$$