In the Name of God

Introduction to Machine Learning (25737-2)

Statistics Review

Spring Semester 1402-03

Department of Electrical Engineering

Sharif University of Technology

Instructor: Dr. R. Amiri

1 Basics

The story of Statistics starts with the following three important theorems;

1.1 Weak Law of Large Numbers

If $X_1, X_2, \dots, X_n$ are n *i.i.d* random variables with $\mathbb{E}[X_i] = \mu$, and $\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$, then for each $\epsilon > 0$, we have:

$$\lim_{n\to\infty} \mathbb{P}\left(|\bar{X}_n - \mu| < \epsilon\right) = 1$$

1.2 Strong Law of Large Numbers

Similarly, we have:
If $X_1, X_2, \dots, X_n$ are n *i.i.d* random variables with $\mathbb{E}[X_i] = \mu$, and $\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$, then we have:

$$\lim_{n\to\infty} \mathbb{P}\left(\bar{X}_n = \mu\right) = 1$$

1.3 Central Limit Theorem

Let $X_1, X_2, \dots, X_n$ be a sequence of *i.i.d* random variables, with mean μ and variance σ^2. Then the distribution of

$$\frac{X_1 + X_2 + \dots + X_n - n\mu}{\sigma\sqrt{n}}$$

tends to the standard normal as $n \to \infty$. That is, for $-\infty < a < \infty$,

$$\mathbb{P}\left\{\frac{X_1 + \dots + X_n - n\mu}{\sigma\sqrt{n} \leq a}\right\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{x^2}{2}} dx \quad as \quad n \to \infty$$

2 Concentration

There are lots of parameters being estimated in Statistics; Mean, Variance, different Moments, even the distribution and … . But how can we make sure this estimations are accurate? Concentration bounds and inequalities are the tools which help us to make sure of our accuracy.

2.1 Markov's Inequality

If X is a non-negative random variable and $a > 0$, then the probability that X is at least a is at most the expectation of X divided by a:

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$$

2.2 Chebyshev's Inequality

Let X be a random variable with finite non-zero variance σ^2 and expected value μ, then for any real number $k > 0$,

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

2.3 Moment Bounding

If ϕ is a non-decreasing non-negative function, X is a random variable, and $\phi(a) > 0$, then:

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(\phi(X)}{\phi(a)}$$

an immediate corollary, using higher moments of X supported on values larger than 0, is

$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}(|X^n|)}{a^n}$$

2.4 Taylor Expansion

Sometimes, it is possible to approximate the moments of a function f of a random variable X using Taylor expansions, provided that f is sufficiently differentiable and that the moments of X are finite. As an example, the *Moment Generating Function* itself can be useful sometimes:

$$M(t) = \mathbb{E}[e^{tX}] = \mathbb{E}[\sum_{n=0}^{\infty} \frac{(tX)^n}{n!}] = \mathbb{E}[\sum_{n=0}^{\infty} X^n \frac{t^n}{n!}] = \sum_{n=0}^{\infty} \frac{t^n}{n!} \mathbb{E}[X^n]$$

2.5 Chernoff's Inequality

Let $X_1, X_2, \ldots, X_n$ be n independent Bernouli random variables with parameters p_i. Consider their sum $S_n = \sum_{i=1}^{n} X_i$ and denote its mean by $\mu = \mathbb{E}[S_n]$. Then, for any $t > \mu$, we have

$$\mathbb{P}\{S_n \geq t\} \leq e^{-\mu} \left(\frac{e\mu}{t}\right)^t .$$

3 Estimators

Estimation, means using all of our given data and information, to make a guess about a true unknown parameter. There are tons of different estimation methods and estimators and each of them are based on certain assumptions and beliefs about the unknown parameters. Note that, it is usually hard to say that which estimator is better than the other, because they all act based on some particular assumptions which we should adapt ourselves to the one which is closest to the reality and the application we are seeking for.

3.1 Bias in Estimation

Statistical bias is a feature of a statistical technique or of its results whereby the expected value of the results differs from the true underlying quantitative parameter being estimated.

$$bias(T, \theta) = bias(T) = \mathbb{E}[T] - \theta$$

3.2 MLE

In statistics, *Maximum Likelihood Estimation* (MLE) is a method of estimating the parameters of an assumed probability distribution, given some observed data. This is achieved by maximizing a likelihood function so that, under the assumed statistical model, the observed data is most probable.

We model a set of observations as a random sample from an unknown joint probability distribution which is expressed in terms of a set of parameters. The goal of maximum likelihood estimation is to determine the parameters for which the observed data have the highest joint probability. We write the parameters governing the joint distribution as a vector $\theta = [\theta_1, \theta_2, \ldots, \theta_k]^\intercal$, so that this distribution falls within a parametric family $\{f(., \theta) | \theta \in \Theta\}$, where Θ is called the *parameter space*, a finite-dimensional subset of Euclidean space. Evaluating the joint density at the observed data sample $y = (y_1, y_2, \ldots, y_n)$ gives a real-valued function,

$$\mathcal{L}_n(\theta) = \mathcal{L}_n(\theta; \mathbf{y}) = f_n(\mathbf{y}, \theta),$$

which is called the likelihood function. For independent and identically distributed random variables, $f_n(\mathbf{y}, \theta)$ will be the product of univariate density functions:

$$f_n(\mathbf{y}, \theta) = \prod_{k=1}^{n} f_k^{univar}(y_k; \theta).$$

The goal of maximum likelihood estimation is to find the values of the model parameters that maximize the likelihood function over the parameter space, that is

$$\hat{\theta} = \arg\max_{\theta \in \theta} \mathcal{L}_n(\theta, \mathbf{y})$$

3.2.1 Unbiased Variance Estimation

Let $x_1, x_2, \ldots, x_n \sim \mathcal{N}(\mu, \sigma^2)$ be independent and identically distributed samples from a normal distribution with mean μ and variance σ^2. The sample mean and unbiased sample variance are given by:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

The resulting t-statistic is given by:

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}} \sim t_{n-1}.$$

and is distributed according to a Student's t distribution with $n - 1$ degrees of freedom.

3.2.2 ML Variance Estimation

Instead of the unbiased estimate s^2 we may also use the maximum likelihood estimate

$$s_{ML}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

yielding the statistic

$$t_{ML} = \frac{\bar{x} - \mu}{\sqrt{\frac{s_{ML}^2}{n}}} = \sqrt{\frac{n}{n-1}} t.$$

This is distributed according to the location-scale t distribution:

$$t_{ML} \sim lst(0, \tau^2 = \frac{n}{n-1}, n-1).$$

3.3 MAP

A *Maximum a Posteriori* probability (MAP) estimate is an estimate of an unknown quantity, that equals the mode of the posterior distribution.Assume that a prior distribution g over θ exists. This allows us to treat θ as a random variable as in Bayesian statistics. We can calculate the posterior distribution of θ using Bayes' theorem:

$$\theta \to f(\theta|x) = \frac{f(x|\theta)g(\theta)}{\int_\theta f(x|\nu)g(\nu)d\nu}$$

where g is density function of θ, Θ is the domain of g.
The method of maximum a posteriori estimation then estimates θ as the mode of the posterior distribution of this random variable:

$$\hat{\theta}_{MAP}(x) = \arg\max_{\theta \in \Theta} f(\theta|x) = \arg\max_{\theta \in \Theta} \frac{f(x|\theta)g(\theta)}{\int_\Theta f(x|\nu)g(\nu)d\nu} = \arg\max_{\theta \in \Theta} f(x|\theta)g(\theta)$$

The denominator of the posterior distribution (so-called marginal likelihood) is always positive and does not depend on θ and therefore plays no role in the optimization.

3.4 Regression

Regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome' or 'response' variable, or a 'label' in machine learning parlance) and one or more independent variables (often called 'predictors', 'covariates', 'explanatory variables' or 'features'). The most common form of regression analysis is linear regression, in which one finds the line (or a more complex linear combination) that most closely fits the data according to a specific mathematical criterion.
In practice, researchers first select a model they would like to estimate and then use their chosen method (e.g., ordinary least squares) to estimate the parameters of that model. Regression models involve the following components:

- The unknown parameters, often denoted as a scalar or vector β.
- The independent variables, which are observed in data and are often denoted as a vector $\mathbf{X}_i$.
- The dependent variable, which are observed in data and often denoted using the scalar $\mathbf{Y}_i$.
- The error terms, which are not directly observed in data and are often denoted using the scalar e_i.

Most regression models propose that $\mathbf{Y}_i$ is a function (regression function) of $\mathbf{X}_i$ and β, with e_i representing an additive error term that may stand in for un-modeled determinants of $\mathbf{Y}_i$ or random statistical noise:

$$\mathbf{Y}_i = f(\mathbf{X}_i, \beta) + e_i$$

The goal is to estimate the function that most closely fits the data.
Once you determine the preferred statistical model, different forms of regression analysis provide tools to estimate the parameters β. For example, least squares (including its most common variant, ordinary least squares) finds the value of β that minimizes the sum of squared errors

$$\sum_i (\mathbf{Y}_i - f(X_i, \beta))^2.$$

4 Decision Theory

4.1 Confidence Interval

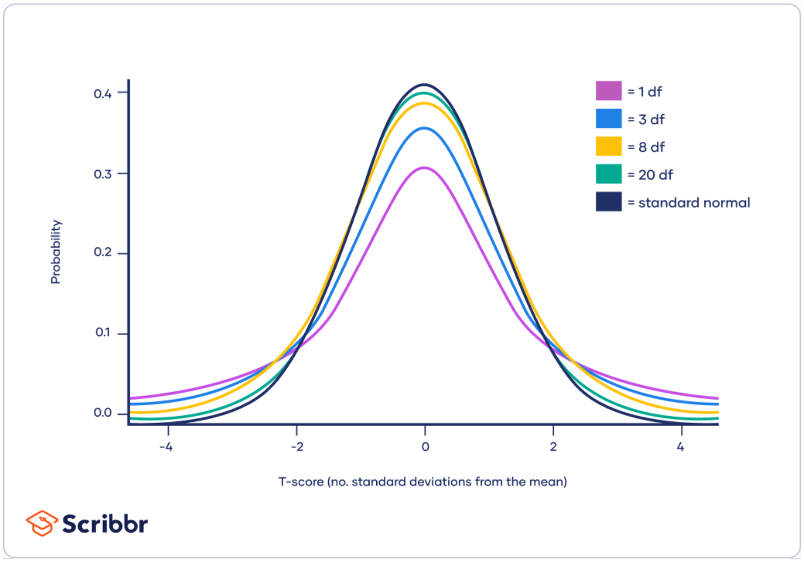
Let X be a random sample from a probability distribution with statistical parameter θ, which is a quantity to be estimated, and ϕ, representing quantities that are not of immediate interest. A confidence interval for the parameter θ, with confidence level or coefficient γ, is an interval $(u(X), v(X))$ determined by random variables $u(X)$ and $v(X)$ with the property:

$$\mathbb{P}\left(u(X) < \theta < v(X)\right) = \gamma \quad \text{for every } (\theta, \phi).$$

4.2 Hypothesis Testing

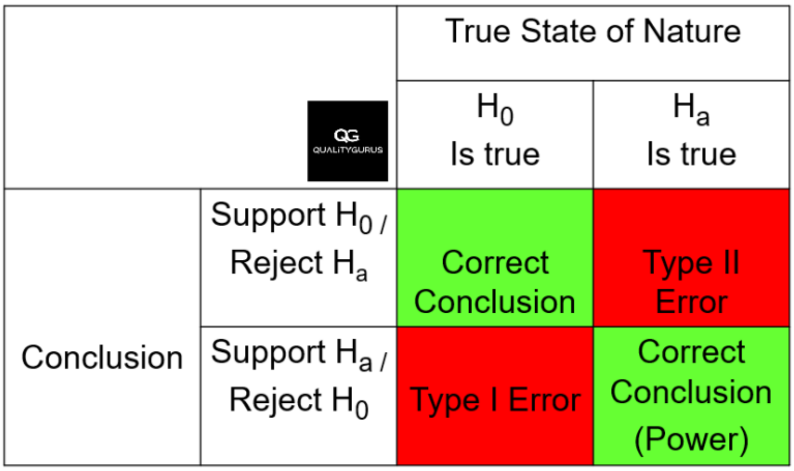
A *Statistical Hypothesis Test* is a method of statistical inference used to decide whether the data sufficiently support a particular hypothesis. A statistical hypothesis test typically involves a calculation of a tests statistic. Then a decision is made, either by computing the test statistic to a critical or equivalently by evaluating a *p-value* computed from the test statistic.

- **Statistical Hypothesis**: A statement about the parameters describing a population(not a sample).
- **Test Statistic**: A value calculated from a sample without any unknown parameters, often to summarized the sample for comparison purposes.
- **Null Hypothesis** (H_0): The statement being tested in a test of statistical significance is called the null hypothesis.
- **Alternative Hypothesis** (H_1): The statement that is being tested against the null hypothesis is the alternative hypothesis.
- **p-value**: the p-value is the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis is correct.



The typical steps involved in performing a hypothesis test in practice are:

1. Define a hypothesis (claim which is testable using data).
2. Select a relevant statistical test with associated test statistic T.
3. Derive the distribution of the test statistic under the null hypothesis from the assumptions. In standard cases this will be a well-known result. For example, the test statistic might follow a Student's t distribution with known degrees of freedom, or a normal distribution with known mean and variance.
4. Select a significance level (α), the maximum acceptable false positive rate. Common values are 5% and 1%.
5. Compute from the observations the observed value t_{obs} of the test statistic T.
6. Decide to either reject the null hypothesis in favor of the alternative or not reject it. The decision rule is to reject the null hypothesis H_0 if the observed value t_{obs} is in the critical region, and not to reject the null hypothesis otherwise.



		True State of Nature	
	QG QUALITYGURUS	H_0 Is true	H_a Is true
Conclusion	Support H_0 / Reject H_a	Correct Conclusion	Type II Error
	Support H_a / Reject H_0	Type I Error	Correct Conclusion (Power)