A Quick Review of Probability and Inferential Statistics for Machine Learning

Masoud Farivar

January 14, 2015

Outline of Part I: Basic Probability

- Definitions
- Random variables: pdf, expectation, variance, typical distributions
- Bounds: Markov, Chebyshev and Chernoff
- Multi-dimensional random variables
- Maximum likelihood estimation
- Central limit theorem

Outline of Part II: Basic Inferential Statistics

- Hypothesis Testing, Confidence intervals
- t-Distribution, t-Test, Paired t-Test
- Wilcoxon Signed-Rank Test

Elements of Probability

Definition:

- Sample Space Ω: Set of all possible outcomes
- Event Space $\mathcal{F}$: A family of subsets of Ω
- Probability Measure: Function $P : \mathcal{F} \to \mathbb{R}$ with properties:
  1. $P(A) \geq 0 \quad (\forall A \in \mathcal{F})$
  2. $P(\Omega) = 1$
  3. A_i's disjoint, then $P(\bigcup_i A_i) = \sum_i P(A_i)$

Conditional Probability and Independence

Conditional probability:

- For events A, B:

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)}$$

- Intuitively means "probability of A when B is known"

Independence

- A, B independent if $P(A|B) = P(A)$ or equivalently: $P(A \bigcap B) = P(A)P(B)$
- Beware of intuition: roll two dies (x_a and x_b), outcomes $\{x_a = 2\}$ and $\{x_a + x_b = k\}$ are independent if $k = 7$, but not otherwise!

Basic laws and bounds

- Union bound: since $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, we have

$$P(\bigcup_i A_i) \leq \sum_i P(A_i)$$

- Law of total probability: if $\bigcup_i A_i = \Omega$, then

$$P(B) = \sum_i P(A_i \cap B) = \sum_i P(A_i)P(B|A_i)$$

- Chain rule: $P(A_1, A_2, \ldots, A_N) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \cdots P(A_N|A_1, \ldots, A_{N-1})$
- Bayes rule: $P(A|B) = P(B|A)\frac{P(A)}{P(B)}$

Random Variables and Distributions

- A random variable X is a function $X : \Omega \to \mathbb{R}$
  Example: Number of heads in 20 tosses of a coin
- Probabilities of events associated with random variables defined based on the original probability function. e.g.,
  $P(X = k) = P(\{\omega \in \Omega | X(\omega) = k\})$
- Cumulative Distribution Function (CDF) $F_X : \mathbb{R} \to [0, 1]$:
  $F_X(x) = P(X \leq x)$
- (X discrete) Probability Mass Function (pmf): $p_X(x) = P(X = x)$
- (X continuous) Probability Density Function (pdf):
  $f_X(x) = dF_X(x)/dx$

Properties of Distribution Functions

- CDF:
  - $0 \leq F_X(x) \leq 1$
  - F_X monotone increasing, with $lim_{x \to -\infty} F_X(x) = 0$, $lim_{x \to \infty} F_X(x) = 1$
- pmf:
  - $0 \leq p_X(x) \leq 1$
  - $\sum_x p_X(x) = 1$
  - $\sum_{x \in A} p_X(x) = p_X(A)$
- pdf:
  - $f_X(x) \geq 0$
  - $\int_{-\infty}^{\infty} f_X(x)dx = 1$
  - $\int_{x \in A} f_X(x)dx = P(X \in A)$

Expectation and Variance

- Assume random variable X has pdf $f_X(x)$, and $g : \mathbb{R} \to \mathbb{R}$. Then

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

- for discrete X, $E[g(X)] = \sum_x g(x) p_X(x)$
- Expectation is linear:
  - for any constant $a \in \mathbb{R}$, $E[a] = a$
  - $E[ag(X)] = aE[g(X)]$
  - $E[g(X) + h(X)] = E[g(X)] + E[h(X)]$
- $Var[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$

Conditional Expectation

- Conditional Expectation (Discrete)

$$E[g(X,Y)|Y=a] = \sum_x g(x,a) p_{X|Y=a}(x)$$

  (similar for continuous random variables)

- Iterated expectation:

$$E[g(X,Y)] = E_a[E[g(X,Y)|Y=a]]$$

Some Common Random Variables

- $X \sim Bernoulli(p)$ $(0 \leq p \leq 1)$: $p_X(x) = \begin{cases} p & \text{x=1,} \\ 1-p & \text{x=0.} \end{cases}$
- $X \sim Geometric(p)$ $(0 \leq p \leq 1)$: $p_X(x) = p(1-p)^{x-1}$
- $X \sim Uniform(a, b)$ $(a < b)$: $f_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases}$
- $X \sim Normal(\mu, \sigma^2)$: $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$
- $X \sim Binomial(n, p)$ $(n > 0, \quad 0 \leq p \leq 1)$:

$$p_X(x) = \left(\begin{array}{c} \text{n} \\ \text{x} \end{array} \right) p^x (1-p)^{n-x}$$

Some Useful Inequalities

- Markov's Inequality: X random variable, and $a > 0$. Then:

$$P(|X| \geq a) \leq \frac{E[|X|]}{a}$$

- Chebyshev's Inequality: If $E[X] = \mu$, $Var(X) = \sigma^2$, $k > 0$, then:

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

- Chernoff bound: Let $X_1, \ldots, X_n$ independent Bernoulli with $P(X_i = 1) = p_i$. Denoting $\mu = E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} p_i$,

$$P(\sum_{i=1}^{n} X_i \geq (1+\delta)\mu) \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}$$

for any δ. Multiple variants of Chernoff-type bounds exist, which can be useful in different settings

Multiple Random Variables and Joint Distributions

$X_1, \ldots, X_n$ random variables

- Joint CDF: $F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n)$
- Joint pdf: $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \frac{\partial^n F_{X_1,\ldots,X_n}(x_1,\ldots,x_n)}{\partial x_1 \ldots \partial x_n}$
- Marginalization: $f_{X_1}(x_1) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) dx_2 \ldots dx_n$
- Conditioning: $f_{X_1|X_2,\ldots,X_n}(x_1|x_2, \ldots, x_n) = \frac{f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)}{f_{X_2,\ldots,X_n}(x_2,\ldots,x_n)}$
- Chain Rule: $f(x_1, \ldots, x_n) = f(x_1) \prod_{i=2}^{n} f(x_i|x_1, \ldots, x_{i-1})$
- Independence: $f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i)$.

Random Vectors

$X_1, \ldots, X_n$ random variables. $X = [X_1 X_2 \ldots X_n]^T$ random vector.

- If $g : \mathbb{R}^n \to \mathbb{R}$, then
  $E[g(X)] = \int_{\mathbb{R}^n} g(x_1, \ldots, x_n) f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) dx_1 \ldots dx_n$
- if $g : \mathbb{R}^n \to \mathbb{R}^m$, $g = [g_1 \ldots g_m]^T$, then
  $E[g(X)] = \left[E[g_1(X)] \ldots E[g_m(X)]\right]^T$
- Covariance Matrix: $\Sigma = Cov(X) = E\left[(X - E[X])(X - E[X])^T\right]$
- Properties of Covariance Matrix:
  - $\Sigma_{ij} = Cov[X_i, X_j] = E\left[(X_i - E[X_i])(X_j - E[X_j])\right]$
  - Σ symmetric, positive semidefinite

Multivariate Gaussian Distribution

$\mu \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{n \times n}$ symmetric, positive semidefinite
$X \sim \mathcal{N}(\mu, \Sigma)$ n-dimensional Gaussian distribution:

$$f_X(x) = \frac{1}{(2\pi)^{n/2} det(\Sigma)^{1/2}} \exp\Big(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\Big)$$

- $E[X] = \mu$
- $Cov(X) = \Sigma$

Parameter Estimation: Maximum Likelihood

- Parametrized distribution $f_X(x;\theta)$ with parameter(s) θ unknown.
- IID samples $x_1, \ldots, x_n$ observed.
- Goal: Estimate θ
- (Ideally) MAP: $\hat{\theta} = argmax_\theta\{f_{\Theta|X}(\theta|X = (x_1, \ldots, x_n))\}$
- (In practice) MLE: $\hat{\theta} = argmax_\theta\{f_{X|\theta}(x_1, \ldots, x_n; \theta)\}$

MLE Example

$X \sim Gaussian(\mu, \sigma^2)$. $\theta = (\mu, \sigma^2)$ unknown. Samples $x_1, \ldots, x_n$. Then:

$$f(x_1, \ldots, x_n; \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}\right)$$

Setting: $\frac{\partial \log f}{\partial \mu} = 0$ and $\frac{\partial \log f}{\partial \sigma} = 0$
Gives:

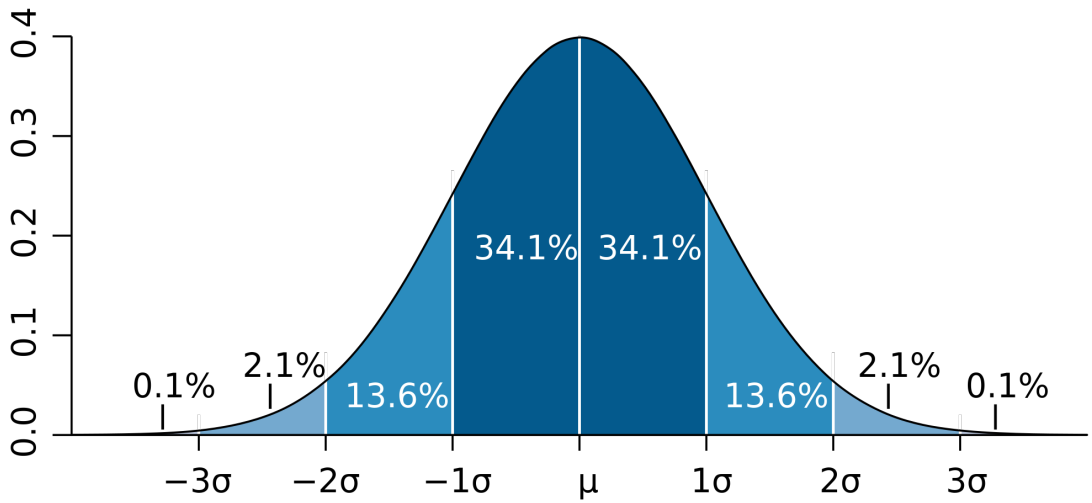
$$\hat{\mu}_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n}, \; \hat{\sigma}^2_{MLE} = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu})^2}{n}$$

Central limit theorem

- Central limit theorem: Let $X_1, X_2, \ldots, X_n$ be iid with finite mean μ and finite variance σ^2, then the random variable $Y = \frac{1}{n}\sum_{i=1}^{n} X_i$ is approximately Gaussian with mean μ and variance $\frac{\sigma^2}{n}$
- Approximation becomes better as n grows
- Law of large numbers as a corollary

Confidence intervals

- Normal Distribution:



- Example: 1.96 is the approximate value of the 97.5 percentile point of the normal distribution. So, we are 95% confident that a sample from a normal distribution lies within $[-1.96\sigma, +1.96\sigma]$.

Hypothesis Testing

The goal of hypothesis testing is generally to rule out chance as a plausible explanation for the results.

- Example: efficacy test for a new drug
  - A sample is selected from the population and the treatment is applied and the results are measured.
  - If the results for the individuals in the sample are noticeably different from the results for the individuals in the original population, we have strong evidence that the treatment has an effect.
  - Otherwise, we can not rule out the possibility that the difference between the sample and the population is simply caused by a sampling error.

Hypothesis Testing

Null Hypothesis (H_0) A maintained hypothesis that is held to be true unless sufficient evidence to the contrary is presented. *importance*.

Alternative Hypothesis (H_1) A hypothesis that is held to be true when the null hypothesis is rejected.

Significance Level (α) The probability of rejecting a true null hypothesis.

***p*-value** The probability of obtaining the observed sample results assuming the null hypothesis is actually true.

Conclusion If the *p*-value is equal or smaller than the significance level (α), it suggests that the observed data are inconsistent with the assumption that the null hypothesis is true, and thus that hypothesis must be rejected and the alternative hypothesis is accepted as true.

Type I / Type II errors

True state of null hypothesis		
Statistical decision	H_0 true	H_0 false
Reject H_0	Type I error α	Correct $1-\beta$
Don't reject H_0	Correct $1-\alpha$	Type II error β

Type I error (false positive)

$$\text{P(Type I error)} = \alpha$$

Type II error(false negative)

$$\text{P(Type II error)} = \beta$$

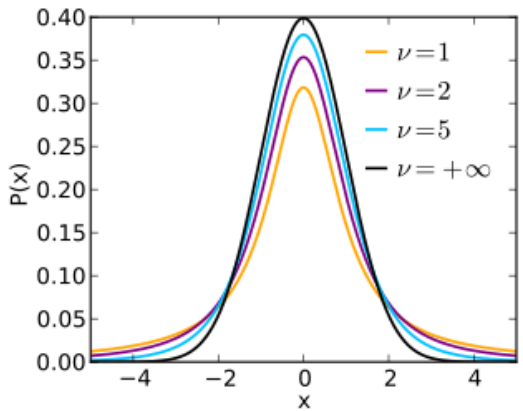
t-distribution (Student's distribution)

- Suppose Z has the standard normal distribution, V has the χ^2 distribution with n degrees of freedom, and that Z and V are independent. Then

$$T = \frac{Z}{\sqrt{V/n}}$$

  has a t-distribution with n degrees of freedom.

- As the degrees of freedom increase, the t-distribution approaches the normal distribution, and is equal to it when $n = \infty$.



t-Test

- A t-test is any statistical hypothesis test in which the test statistic follows a Student's t distribution if the null hypothesis is supported.
- One-Sample t-Test:
  - Define Null and Alternative Hypotheses
  - Choose α
  - Calculate degrees of freedom
  - State the decision rule
  - Calculate test statistic
  - State the conclusion
- Paired t-Test (dependent samples):
  - Used when comparing the performance of two models/methods
  - Calculate the difference score for each pairing (loss on each test point)
  - run a one-sample t-test

Wilcoxon Signed-Rank Test

1. It can be used as an alternative to the paired t-test
2. Compute the difference for each data pair, drop zeros from the list
3. Order the absolute differences (ignoring signs) and rank them (replace tied values with the average of ranks)
4. The *signed-rank statistic* is $S =$ the sum of the ranks for the pairs with a positive difference. Compute S for this data.

$$\text{Mean}(S) = n(n+1)/4 \qquad \text{SD}(S) = \sqrt{\sum_{i=1}^{n} R_i^2/4},$$

where n is the number of non-zero differences and $\sum_{i=1}^{n} R_i^2$ is the sum of the squares of all the ranks.

5. If n is not too small (and no excessive number of ties), then

$$Z = \frac{S - \text{Mean}(S)}{\text{SD}(S)} = \frac{S - n(n+1)/4}{\sqrt{\sum_{i=1}^{n} R_i^2/4}}$$

should be approximately normal.

Example of Wilcoxon Signed-Rank Test

A psychologist wants to test the hypothesis that alcohol consumption favors Boundary extension (the tendency to remember scenes as if they included information beyond the boundaries).

Example

```
Alc=[6/10 4/10 5/10 6/10 3/10 3/10 6/10 7/10 8/10 2/10];
NoAlc=[1/10 3/10 3/10 6/10 3/10 2/10 5/10 6/10 6/10 3/10];
[p,h,stats]= signrank(Alc , NoAlc);
```

This Matlab code outputs h =1 and p =0.0391. Hence, the Wilcoxon signed rank test indicates that we can reject the null hypothesis at level 0.05 of significance: alcohol consumption affects boundary extension occurances.

References

The source of this review are the following:

- Course notes from CMU's 10-701
- Course notes from Stanford's CS224w
- Borgo, Mauro, Alessandro Soranzo, and Massimo Grassi. "MATLAB for Psychologists", Springer, 2012.
- Wikipedia!