# Machine Learning

CSE 6363 (Fall 2019)

## Lecture 2 Variance, STD, t-Test and correlation

Dajiang Zhu, Ph.D.

Department of Computer Science and Engineering

***Most slides of this lecture courtesy: Dr. Heng Huang***

# Deviation

- Deviation: the distance of each value from the mean. If the mean is 3, a value of 4 has a deviation of 1 (subtract the mean from the value).
- Deviation can be positive or negative.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| x: | 6 | 2 | 0 | 0 | 1 | 3 | $\bar{x}$: 2 |
| X-$\bar{x}$ : | 4 | 0 | -2 | -2 | -1 | 1 | |

# Standard Deviation

## Standard Deviation

The Standard Deviation is a measure of how spread out numbers are.

Its symbol is **σ** (the greek letter sigma)

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}.$$

- It is a special form of average deviation from the mean.
- It is affected by the individual items in the distribution
- It is considered as the most reliable measure of variability.
- N-1?

# Variance (first impression)

**Square of Standard Deviation**

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

**Sampled Variance?**

# Expectation

The expectation of a random variable X is written as $\mathbb{E}(X)$

Let $X$ be a **continuous** random variable

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\, dx$$

Let $X$ be a **discrete** random variable

$$\mathbb{E}(X) = \sum_x x f_X(x) = \sum_x x \mathbb{P}(X = x)$$

$$\mathbf{E}[\boldsymbol{X} + \boldsymbol{Y}] = \mathbf{E}[\boldsymbol{X}] + \mathbf{E}[\boldsymbol{Y}]$$
$$\mathbf{E}[\boldsymbol{a}\boldsymbol{X}] = \boldsymbol{a}\,\mathbf{E}[\boldsymbol{X}]$$

# Variance

The **Variance** of a random variable **X** is the **Expectation** of the squared **Deviation** from the **Mean of X**

$$\operatorname{Var}(X) = \operatorname{E}\left[(X-\mu)^2\right] = \operatorname{E}\left[X^2\right] - \operatorname{E}[X]^2$$

For **Discrete** random variable:

$$\operatorname{Var}(X) = \sum_{i=1}^{n} p_i \cdot (x_i - \mu)^2 \qquad \mu = \sum_{i=1}^{n} p_i \cdot x_i$$

$$\operatorname{Var}(X) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2$$

$$\operatorname{Var}(X+a) = \operatorname{Var}(X) \qquad \operatorname{Var}(aX) = a^2 \operatorname{Var}(X)$$

# Covariance

- The **Covariance** between two jointly distributed real-valued random variables **X** and **Y** is defined as the expected product of their deviations from their individual expectations:

$$\operatorname{cov}(X,Y) = \operatorname{E}\left[(X - \operatorname{E}[X])(Y - \operatorname{E}[Y])\right]$$

$$\begin{aligned}
\operatorname{cov}(X,Y) &= \operatorname{E}[(X - \operatorname{E}[X])\,(Y - \operatorname{E}[Y])] \\
&= \operatorname{E}[XY - X\operatorname{E}[Y] - \operatorname{E}[X]Y + \operatorname{E}[X]\operatorname{E}[Y]] \\
&= \operatorname{E}[XY] - \operatorname{E}[X]\operatorname{E}[Y] - \operatorname{E}[X]\operatorname{E}[Y] + \operatorname{E}[X]\operatorname{E}[Y] \\
&= \operatorname{E}[XY] - \operatorname{E}[X]\operatorname{E}[Y].
\end{aligned}$$

- For random vectors $\mathbf{X} \in \mathbb{R}^m$ and $\mathbf{Y} \in \mathbb{R}^n$, the mxn covariance matrix is:

$$\begin{aligned}
\operatorname{cov}(\mathbf{X},\mathbf{Y}) &= \operatorname{E}\left[(\mathbf{X} - \operatorname{E}[\mathbf{X}])(\mathbf{Y} - \operatorname{E}[\mathbf{Y}])^{\mathrm{T}}\right] \\
&= \operatorname{E}\left[\mathbf{X}\mathbf{Y}^{\mathrm{T}}\right] - \operatorname{E}[\mathbf{X}]\operatorname{E}[\mathbf{Y}]^{\mathrm{T}}
\end{aligned}$$

# Covariance

- **Covariance** is a measure of the association or dependence between two random variables X and Y. Covariance can be either positive or negative. (Variance is always positive)

- cov(X, Y ) will be **positive** if large values of X tend to occur with large values of Y , and small values of X tend to occur with small values of Y . For example, if X is height and Y is weight of a randomly selected person, we would expect cov(X, Y ) to be positive.

- cov(X, Y ) will be **negative** if large values of X tend to occur with small values of Y , and small values of X tend to occur with large values of Y . For example, if X is age of a randomly selected person, and Y is heart rate, we would expect X and Y to be negatively correlated (older people have slower heart rates).

- If X and Y are **independent**, then there is no pattern between large values of X and large values of Y , so **cov(X, Y ) = 0**. However, cov(X, Y ) = 0 does NOT imply that X and Y are independent.

# Correlation

- The **correlation** between X and Y , also called the **Correlation Coefficient (Pearson's correlation)**, is given by

$$\operatorname{corr}(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}}$$

  - The correlation measures linear association between X and Y. It takes values only between -1 and +1, and has the same sign as the covariance.
  - The correlation is $\pm 1$ if and only if there is a perfect linear relationship between X and Y, i.e. corr(X, Y ) = 1 $\Longleftrightarrow$ Y = aX + b for some constants a and b.
  - **The correlation is 0 if X and Y are independent, but a correlation of 0 does not imply that X and Y are independent.**

# t-Test

- One-sample: if the mean of a population has a value specified in the null hypothesis (right tailed, left tailed and two tailed).

- Two-sample: to test the null hypothesis that the means of two populations are equal. (paired and unpaired)

- Example:

Two sample and assume equal variances, the test statistic is calculated as:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$s^2 = \frac{\sum_{i=1}^{n_1}(x_i - \bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_j - \bar{x}_2)^2}{n_1 + n_2 - 2}$$