COMP4121 Advanced Algorithms

Aleks Ignjatović

School of Computer Science and Engineering
University of New South Wales

Some Introductory Notes on Statistics

A few basic fact from Statistics

- **A random vector** is a sequence of n random variables $(X_1, \ldots, X_n)$.

- We denote the (multivariate) PDF of such a vector by $f_{\mathbf{X}}(x_1, \ldots, x_n)$.

- Random variables $X_1, \ldots, X_n$ are **independent** if for all subsets $A_1, \ldots, A_i \subset \mathbb{R}$,

$$\mathcal{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} \mathcal{P}(X_i \in A_i). \tag{1}$$

- This happens just in case

$$f_{\mathbf{X}}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i). \tag{2}$$

- If random variables $X_1, \ldots, X_n$ are independent and have the same probability distribution F, we say that $X_1, \ldots, X_n$ are **IID (independent identically distributed)** random variables, or **random sample of size n from** F.

- The **sample mean** is the random variable defined by $\overline{X} = \dfrac{X_1 + \ldots + X_n}{n}$.

- Assume $X_1, \ldots, X_n$ are IID with mean $E(X_i) = \mu$ and variance $V(X_i) = \sigma^2$; then, since X_i are equally distributed, the expected value of $\overline{X}$ and the variance $V(\overline{X})$ satisfy

$$E(\overline{X}) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu \tag{3}$$

A few basic fact from Statistics

- Using this we get

$$V(\overline{X}) = E(\overline{X} - E(\overline{X}))^2 = E(\overline{X} - \mu)^2 = E\left(\frac{\sum_{i=1}^{n}(X_i - \mu)}{n}\right)^2 =$$

$$E\left(\frac{1}{n^2}\sum_{i=1}^{n}(X_i - \mu)^2 + \frac{2}{n^2}\sum_{\substack{i \neq j}}^{n}(X_i - \mu)(X_j - \mu)\right) =$$

$$\frac{1}{n^2}\left(\sum_{i=1}^{n}E(X_i - \mu)^2 + 2\sum_{\substack{i \neq j}}^{n}E((X_i - \mu)(X_j - \mu))\right);$$

- We now use the fact that if X_i, X_j are independent then their covariance is 0:

$$E((X_i - \mu)(X_j - \mu)) = E(X_i - \mu)E(X_j - \mu) = 0 \times 0 = 0$$

- Thus

$$V(\overline{X}) = \frac{1}{n^2} \sum_{i=1} E(X_i - \mu)^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}. \tag{4}$$

- Note that this explains why taking the mean of multiple measurements provides more accurate value than a single measurement: if the variance of a single measurement is σ^2 then the variance of the mean of n measurements is only σ^2/n.

- Assume now that we have n measurements X_i of a quantity using the same instrument and w[illegible]imate the variance of the instrument.

- Since the expected va[illegible] is μ one might think that

$$E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\right)\stackrel{??}{=}E\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right)=\frac{1}{n}\sum_{i=1}^{n}E(X_i-\mu)^2=\frac{n\,\sigma^2}{n}=\sigma^2$$

- This would make $\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$ an unbiased estimator for the variance of each of X_i, but this is not quite so - the "equality" with the question marks fails, because we do not take into account that in fact $\overline{X}$ is not quite equal to μ, making the expected value of the lefthand side slightly smaller than σ^2 because $\eta=\overline{X}$ in fact minimises the value of $s(\eta)=\sum_{i=1}^{n}(X_i-\eta)^2$.

- An unbiased estimate of the variance of all X_i is given by the **sample variance** S_n^2 defined as

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2. \tag{5}$$

To see that the expected value of the sample variance is equal to the variance σ^2 of all X_i, note that

$$E\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = E\left(\sum_{i=1}^{n}(X_i^2 - 2X_i\overline{X} + \overline{X}^2)\right) = \tag{6}$$

$$E\left(\sum_{i=1}^{n}X_i^2 - 2\sum_{i=1}^{n}X_i\overline{X} + \sum_{i=1}^{n}\overline{X}^2\right) = E\left(\sum_{i=1}^{n}X_i^2 - 2n\overline{X}^2 + n\overline{X}^2\right) =$$

$$E\left(\sum_{i=1}^{n}X_i^2 - n\overline{X}^2\right) = \sum_{i=1}^{n}E(X_i^2) - n\,E(\overline{X}^2) = nE(X_1^2) - n\,E(\overline{X}^2) \tag{7}$$

A few basic fact from Statistics

- We now use the fact that for every random variable Y with mean μ and standard deviation σ we have

$$\begin{aligned}\sigma^2 =& E(Y-\mu)^2 = E(Y^2 - 2\mu Y + \mu^2) = E(Y^2) - 2\mu E(Y) + \mu^2 = \\ & E(Y^2) - 2\mu^2 + \mu^2 = E(Y^2) - \mu^2 \end{aligned} \tag{8}$$

Thus, $E(Y^2) = \sigma^2 + \mu^2$. Continuing (7) and using (3) and (4)

$$\begin{aligned} E\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right) &= n E(X_1^2) - n\, E(\overline{X}^2) = n(\sigma^2 + \mu^2) - n\, E(\overline{X}^2) \\ &= n(\sigma^2 + \mu^2) - n(E(\overline{X} - \mu)^2 + \mu^2) = \\ &= n(\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right) = \\ &= (n-1)\sigma^2 \end{aligned} \tag{9}$$

- This clearly implies that $E(S_n^2) = E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = \sigma^2$, i.e., that S_n^2 is an unbiased estimator for the variance of X.

A tool we will often need: Markov Inequality

- Assume $X > 0$ is a non-negative random variable;
- assume also that $t > 0$ is any positive real number;

then:
$$P\{X \geq t\} \leq \frac{E[X]}{t}$$

- **Proof:** Essentially, into account only events when $X \geq t$ and ignore events when $X < t$:
- If X is discrete, then

$$\begin{aligned} E[X] &= \sum_{v} P\{X = v\} \cdot v = \sum_{0 \leq v < t} P\{X = v\} \cdot v + \sum_{v \geq t} P\{X = v\} \cdot v \\ &\geq \sum_{v \geq t} P\{X = v\} \cdot v \geq \sum_{v \geq t} P\{X = v\} \cdot t = t \sum_{v \geq t} P\{X = v\} \\ &= t\, P\{X \geq t\} \end{aligned}$$

- Divide now both sides by $t > 0$ to obtain the Markov Inequality.
- The case when X is continuous is essentially identical and you can find it in the probability refresher lecture notes.

Application of Markov's Inequality: Chebyshev's Inequality (Markov was a student of Chebyshev)

- **Theorem:** Let X be a random variable with a finite expectation $E[X]$ and a finite variance $V[X]$. Then

$$\mathcal{P}(|X - E[X]| \geq t) \leq \frac{V[X]}{t^2}$$

- **Proof:** Note that

$$\mathcal{P}(|X - E[X]| \geq t) = \mathcal{P}((X - E[X])^2 \geq t^2)$$

- Also, $Y = (X - E[X])^2$ is a non-negative random variable with a finite expectation $E[Y] = E((X - E[X])^2) = V[X]$.
- Thus, we can apply the Markov inequality to Y to obtain

$$\mathcal{P}(|X - E[X]| \geq t) = \mathcal{P}((X - E[X])^2 \geq t^2) = \mathcal{P}(Y > t^2) \leq \frac{E[Y]}{t^2} = \frac{V[X]}{t^2}$$

Law of Large Numbers

- If we get a large number of independent samples $X_1, \ldots, X_n$ of a random variable X with an expectation $E[X]$ we would expect that the mean of these samples should be close to the expected value $E[X]$ of X. The Law of large numbers bounds the probability that such a mean is further away from $E[X]$ than a prescribed value ε.

$$\mathcal{P}\left(\left|\frac{X_1 + \ldots + X_n}{n} - E[X]\right| > \varepsilon\right) \leq \frac{V[X]}{n\varepsilon^2}$$

- **Proof:** By the Chebyshev inequality,

$$\mathcal{P}\left(\left|\frac{X_1 + \ldots + X_n}{n} - E\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right]\right| > \varepsilon\right) \leq \frac{V\left[\frac{X_1+X_2+\ldots+X_n}{n}\right]}{\varepsilon^2}$$

- We now use the fact that the expectation is a linear operator and all X_i are equally distributed to conclude that

$$E\left[\frac{X_1 + X_2 + \ldots + X_n}{n}\right] = \frac{E[X_1] + E[X_2] + \ldots + E[X_n]}{n} = \frac{n\,E[X]}{n} = E[X]$$

- On the other hand, variance of any random variable X satisfies $V[a\,X] = a^2 V[X]$.
- Also, for INDEPENDENT variables $X_1, \ldots, X_n$ we have $V[X_1 + \ldots + X_2] = V[X_1] + \ldots + V[X_n]$.

- Thus

$$\mathcal{P}\left(\left|\frac{X_1+\ldots+X_n}{n} - E\left[\frac{X_1+\ldots+X_n}{n}\right]\right| > \varepsilon\right) \leq \frac{V\left[\frac{X_1+\ldots+X_n}{n}\right]}{\varepsilon^2}$$

  implies

$$\mathcal{P}\left(\left|\frac{X_1+\ldots+X_n}{n} - E[X]\right| > \varepsilon\right) \leq \frac{\frac{n\,V[X]}{n^2}}{\varepsilon^2} = \frac{V[X]}{n\,\varepsilon^2}$$

- To summarise, if $X_1, \ldots, X_n$ are independent, equally distributed random variables with a finite expectation $E[X_i] = \mu$ and variance $V[X_i] = v$, then

$$E\left[\frac{X_1+\ldots+X_n}{n}\right] = \mu;$$

$$V\left[\frac{X_1+\ldots+X_n}{n}\right] = \frac{v}{n}$$

$$\mathcal{P}\left(\left|\frac{X_1+\ldots+X_n}{n} - \mu\right| > \varepsilon\right) \leq \frac{v}{n\epsilon^2}$$

- The likelihood of a set of parameter values, θ, given outcomes $\mathbf{x}$, is equal to the probability of those observed outcomes given those parameter values. The likelihood function is defined differently for discrete and continuous probability distributions.

- **Likelihood in case of a discrete probability distribution.** Let X be a random variable with a discrete probability distribution p depending on a parameter θ. Then the function

$$\mathcal{L}(\theta|x) = p_\theta(x) = P_\theta(X = x),$$

considered as a function of θ, is called the likelihood function of θ, given the outcome x of the random variable X.

Maximum Likelihood Estimation

- **Likelihood in case [illegible]tinuous probability distribution.** Let X be a random var[illegible] a continuous probability distribution with density function f[illegible]g on a parameter θ. Then the function

$$\mathcal{L}(\theta|x) = f_\theta(x),$$

considered as a function of θ, is called the likelihood function of θ, given the outcome x of X.

- In case of several random variables, the likelihood function is equal to the joint probability density of the vector $\mathbf{X} = (X_1, \cdots, X_n)$, but seen as a function of unknown parameters of the distribution of X_i with the values of X_i treated as constants.

Maximum Likelihood Estimation

- Thus, if X_i are IID, the likelihood is just the product of the values of the density function at the corresponding values X_i:

$$\mathcal{L}_n(\theta|\mathbf{X}) = \prod_{i=1}^{n} f_\theta(X_i)$$

- To make dependence on the unknown parameters explicit we instead write

$$\mathcal{L}_n(\theta|\mathbf{X}) = \prod_{i=1}^{n} f(X_i;\theta)$$

Maximum Likelihood Estimation

- One reasonable way to choose the values of the unknown parameters is to choose the values which *maximise the likelihood function*, with the intuition behind that we choose the values of the parameters for which the outcome we have is most likely to happen.
- Note that likelihood is NOT the same as probability; in fact it has a somewhat "reverse" role in the following sense:

  - if we have a coin with the probability of getting a head equal to p, than we can calculate the *probability* to get a particular outcome if we toss it 16 times, say the probability that we will get $HHTHTHHHHTTHHTHHT$.

  - if we have already performed the experiment with a coin we know nothing about and observed such an outcome, we can now ask the "reverse question": for what p is the probability of the observed outcome maximal, i.e., for what value of p is such outcome *most likely*?

- Note that if we have gotten 10 heads it is intuitively clear that such outcome is most likely t[illegible]me from a coin for which the probability p of getting [illegible] $\frac{10}{16}$.

- To verify such a hypothesis, we compute the probability of such an outcome, to get 10 heads in that particular order, as a function of p:

$$\mathbb{P}(p) = p^{10}(1-p)^6.$$

To find when such probability is the largest, we look for the stationary points of $\mathbb{P}(p)$, i.e., for p such that $\frac{\partial \mathbb{P}}{\partial p} = 0$. Since

$$\begin{aligned}\frac{\partial \mathbb{P}}{\partial p} &= 10p^9(1-p)^6 + p^{10} \cdot 6(1-p)^5(-1) \\ &= p^9(1-p)^5(10(1-p) - 6p) \\ &= p^9(1-p)^5(10-16p),\end{aligned}$$

$\mathbb{P}(p)$ has a maximum value for $16p = 10$, i.e., for $p = \frac{10}{16}$.

- A ML estimate is usually good for large number of samples, but it can perform really badly if we have only a small number of samples. As an example, consider the following problem.

- Assume that I have given you a box which contains n balls which are numbered consecutively 1 to n, but I do not tell you what n is, i.e., how many balls there are inside. You are allowed to draw one single ball and look at its number, and then you have to estimate how many balls there are inside.

Assume that you drew the ball numbered k. Since all balls are equally likely, if there are n balls inside, then the probability of drawing any particular ball is $1/n$. Thus, the event that you drew ball k has the highest possible probability if $1/n$ is as large as possible, i.e., when n is as small as possible. Since you know that there are at least k balls inside, the MLE estimate for the number of balls in the box is $n = k$, i.e., the MLE estimator in this case is $N(X) = X$.

- What is the mean of this estimator?

- The expected value of this estimator $\mu = E(X)$ is then given by

$$\mu = \sum_{i=1}^{n} \left(\frac{i}{n} \right) = \frac{n(n+1)}{2n} = \frac{n+1}{2}.$$

- Thus, in this case the MLE estimator is extremely biased, because its expected value is only about one half of the true value n of the number of balls inside the box!

- If you instead use the estimator $Y(X) = 2X - 1$, then the expected value of Y is

$$\sum_{i=1}^{n} \frac{2i-1}{n} = \frac{2\sum_{i=1}^{n} i}{n} - \frac{\sum_{i=1}^{n} 1}{n} = \frac{2n(n+1)}{2n} - 1 = n$$

and so this estimator is unbiased – much better than the MLE.

- This does not happen for large samples, it can be shown that as the size of the sample increases, ML estimate approaches the best possible estimate.

- More precisely, while Maximum-Likelihood estimators have no optimum properties for finite samples, in the sense that (when evaluated on finite samples) other estimators may have greater concentration around the true parameter-value, maximum likelihood estimation possesses a number of attractive limiting properties: as the sample size increases to infinity, sequences of maximum likelihood estimators have these properties:
- **Consistency:** the sequence of MLEs converges in probability to the value being estimated.

- **Efficiency:** ML achieves the CramrRao lower bound when the sample size tends to infinity. This means that no consistent estimator has lower asymptotic variance than the MLE.

An example of a Maximum Likelihood Estimation

- Assume that we have n sensors such that the errors of readings $X_1, \ldots, X_n$ of these sensors are independent, unbiased and normally distributed with different and **known** standard deviations $\sigma_1, \ldots, \sigma_n$ (for example, we have tested them all in a lab doing many measurements and comparing their readings with the true value obtained from a precise instrument).
- Assume that, using these sensors, we have obtained measurements $X_1, \ldots, X_n$, which we would like to use to estimate the true value of the quantity being measured.
- Let the true value of the measured quantity be equal to μ.
- Since the sensors are unbiased the expected value of the readings of all sensors is equal to μ, i.e., $E[X_i] = \mu$.
- We can now compute the **likelihood** that a particular μ produced measurements $X_1, X_2, \ldots, X_n$ and then pick the value of μ for which such a likelihood is the largest.

- Since the errors are independent and normally distributed and since $E[X_i] = \mu$, the probability to obtain readings $\mathbf{X} = X_1, \ldots, X_n$ is

$$\mathcal{L}_n(\mu|\mathbf{X}) = \prod_{i=1}^{n} \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(X_i-\mu)^2}{\sigma_i^2}} = \left(\prod_{i=1}^{n} \frac{1}{\sigma_i\sqrt{2\pi}}\right) e^{-\frac{1}{2}\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma_i^2}}$$

- Differentiating with respect to μ and setting the derivative equal to zero we get

$$\frac{d}{d\mu}\mathcal{L}_n(\mu|\mathbf{X}) = \left(\prod_{i=1}^{n} \frac{1}{\sigma_i\sqrt{2\pi}}\right) e^{-\frac{1}{2}\sum_{i=1}^{n}\frac{(X_i-\mu)^2}{\sigma_i^2}} \sum_{i=1}^{n} \frac{(X_i-\mu)}{\sigma_i^2}$$

- Thus,

$$\frac{d}{d\mu}\mathcal{L}_n(\mu|\mathbf{X}) = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{n}\frac{X_i}{\sigma_i^2} - \mu\sum_{i=1}^{n}\frac{1}{\sigma_i^2} = 0 \quad \Leftrightarrow \quad \mu = \frac{\sum_{i=1}^{n}\frac{X_i}{\sigma_i^2}}{\sum_{j=1}^{n}\frac{1}{\sigma_j^2}}$$

An example of a Maximum Likelihood Estimation

- Since the Maximum Likelihood estimate of the true value is

$$\mu = \frac{\sum_{i=1}^{n} \frac{X_i}{\sigma_i^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_i^2}} = \sum_{i=1}^{n} \frac{\frac{1}{\sigma_i^2}}{\sum_{j=1}^{n} \frac{1}{\sigma_j^2}} X_i$$

  we see that the maximum likelihood estimate is a weighted mean of the readings of **all** sensors where each sensor's reading is weighted inversely proportionally to its variance.

- Let us find the variance of such an estimator:

$$V(M(\mathbf{X})) = E\left(\sum_{i=1}^{n} \frac{\frac{1}{\sigma_i^2}}{\sum_{j=1}^{n} \frac{1}{\sigma_j^2}} X_i - \mu\right)^2 = E\left(\sum_{i=1}^{n} \frac{\frac{1}{\sigma_i^2}(X_i - \mu)}{\sum_{k=1}^{n} \frac{1}{\sigma_k^2}}\right)^2$$

$$= E\left(\sum_{i,j=1}^{n} \frac{\frac{1}{\sigma_i^2}(X_i - \mu)}{\sum_{k=1}^{n} \frac{1}{\sigma_k^2}} \frac{\frac{1}{\sigma_j^2}(X_j - \mu)}{\sum_{k=1}^{n} \frac{1}{\sigma_k^2}}\right)$$

An example of a Maximum Likelihood Estimation

- Since the errors of the measurements are pairwise independent, we have $E((X_i - \mu)(X_j - \mu)) = 0$ for $i \neq j$; thus, from

$$V(M(\mathbf{X})) = E\left(\sum_{i,j=1}^{n} \frac{\frac{1}{\sigma_i^2}(X_i - \mu)}{\sum_{k=1}^{n} \frac{1}{\sigma_k^2}} \frac{\frac{1}{\sigma_j^2}(X_j - \mu)}{\sum_{k=1}^{n} \frac{1}{\sigma_k^2}}\right)$$

we obtain

$$V(M(\mathbf{X})) = E\left(\sum_{i=1}^{n} \frac{\frac{1}{\sigma_i^4}(X_i - \mu)^2}{\left(\sum_{j=1}^{n} \frac{1}{\sigma_j^2}\right)^2}\right) = \sum_{i=1}^{n} \frac{\frac{1}{\sigma_i^4}\sigma_i^2}{\left(\sum_{j=1}^{n} \frac{1}{\sigma_j^2}\right)^2} = \frac{1}{\sum_{j=1}^{n} \frac{1}{\sigma_j^2}}.$$

An example of a Maximum Likelihood Estimation

- Since for all i

$$V(M(\mathbf{X})) = \frac{1}{\sum_{j=1}^{n} \frac{1}{\sigma_j^2}} < \frac{1}{\frac{1}{\sigma_i^2}} = \sigma_i^2$$

we get that also $V(M(\mathbf{X})) < \min_{1 \leq i \leq n} \sigma_i^2$.

- Thus, we see that see that our ML estimate is more accurate than the best of our sensors!
- Can we find a more accurate estimator which is also unbiased?
- It turns out that in this case the ML estimator achieves the smallest possible variance among all unbiased estimators.
- Such a lower bound for the variances of all unbiased estimators is given by the Cramer - Rao Theorem which you can look up in the lecture notes on Iterative Filtering which you can find at the class website, look at page 6 of `https://www.cse.unsw.edu.au/~cs4121/lectures_2019/IF.pdf`