CSE 840: Computational Foundations of Artificial Intelligence November 20, 2023

Expectation, Covariance, Some Inequalities and Distributions

Instructor: Vishnu Boddeti *Scribe: Yuping Lin, Hanbing Wang, Weikang Ding*

1 Expectation and Variance in the General Setting

Definition 1 *Define $L^k(\Omega, \mathcal{A}, P)$ space as:*

$$L^k(\Omega, \mathcal{A}, P) := \{X : \Omega \to \mathbb{R} | X \textit{ measurable and } \int_\Omega |X|^k \mathrm{d}P < \infty\}$$

(Here, "$\int_\Omega |X|^k \mathrm{d}P < \infty$" means that the integral $\int_\Omega |X|^k \mathrm{d}P$ exists.)

An $L^k(\Omega, \mathcal{A}, P)$ space is the set of all functions $X : \Omega \to \mathbb{R}$ that are measurable. $(\Omega, \mathcal{A}, P)$ denotes a probability space, where Ω is the sample space, $\mathcal{A}$ is the σ-algebra, and $P_X = X(P)$ is the probability distribution.

Definition 2 *If X is once-integrable, that is, $x \in L^1(\Omega, \mathcal{A}, P)$, the expectation of X is defined as:*

$$E(X) := \int_\Omega X \mathrm{d}P = \int_\mathbb{R} x \mathrm{d}P_X(x)$$

In case that P_X is the probability density, $E(X) := \int_\mathbb{R} x f(x) \mathrm{d}x$. It is also called the first moment of X.

Similarly, if $X^k \in L^1(\Omega, \mathcal{A}, P)$, then

$$E(X^k) = \int X^k \mathrm{d}P$$

is called the k-th moment of X.

If $X^k \in L^2(\Omega, \mathcal{A}, P)$, we define

$$\mathrm{Var}(x) = E((x - E(x))^2)$$
$$\mathrm{Cov}(x, y) = E((x - E(x) \cdot (y - E(y)))$$

2 Markov and Chebyshev Inequalities

2.1 Cauchy-Schwatz Inequality

Theorem 3 *<u>Cauchy-Schwatz Inequality.</u> Let $x, y \in L^2(\Omega, \mathcal{A}, P)$. Then,*

$$E(x \cdot y)^2 \leq E(x^2) \cdot E(y^2)$$

2.2 Markov Inequality

Theorem 4 *Markov Inequality. For* $\forall \varepsilon > 0, f : [0, \infty) \to [0, \infty)$, *if* f *is a monotonically increasing function, then*

$$P(|y| > \varepsilon) \leq \frac{E(f(|y|))}{f(\varepsilon)}$$

In particular, take a special case of $f(x) = x$,

$$P(|y| > \varepsilon) \leq \frac{E(|y|)}{\varepsilon}$$

2.3 Chebyshev Inequality

Theorem 5 *Chebyshev Inequality. For* $\forall \varepsilon > 0, x \in L^2(\Omega, \mathcal{A}, P)$, *we have:*

$$P(|x - E(x)| > \varepsilon) \leq \frac{\mathrm{Var}(x)}{\varepsilon^2}$$

Note that Theorem 5 proves that the probability $P(|x - E(x)| > \varepsilon)$ is loosely (if ε is small) bounded by $\frac{\mathrm{Var}(x)}{\varepsilon^2}$ with no other assumptions. This is an important quantity in learning theory.

3 Examples of Probability Distributions

Discrete distributions:

Definition 6 *Uniform distribution on {1,...,n}*

$$P(\{i\}) = \frac{1}{n}$$

Definition 7 *Binomial distribution on {0,...,n}*
Toss a coin n *times, independently, each time with probability* p *of observing head.*
Denote head=1, tail=0, $x :=$ *# heads.*

$$P(X = k) := \binom{n}{k} p^k (1-p)^{n-k}$$

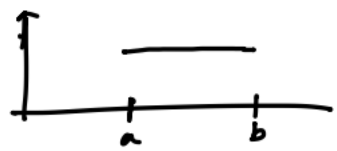
Definition 8 *Poisson distribution on N*

Parameter $\lambda > 0$

$$P(X = k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}$$

Intuition: number of jobs submitted to a cloud service.

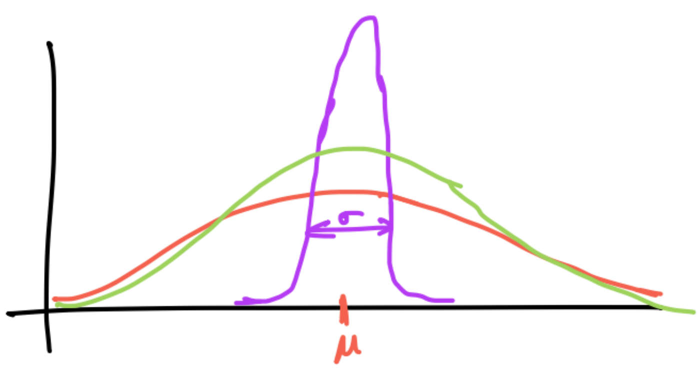
Definition 9 *Continuous distribution:*
Uniform distribution on $[a, b]$*: constant density*



4 Normal Distribution on R

Definition 10 *Density: parameter* μ *(mean),* σ *(std deviation)*

$$f_{\mu,\sigma}(x) := \frac{1}{\sqrt{2\pi\sigma^2}} exp(-\frac{(x-u)^2}{2\sigma^2})$$



Notation: $N(\mu, \sigma^2)$
Some properties:
$x \sim N(\mu_1, \sigma_1^2)$, $y \sim N(\mu_2, \sigma_2^2)$
x, y are independent
then $x + y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_1^2)$

5 Normal distribution in higher dimensions

$X : \Omega \to \mathbb{R}^n$, $X = \begin{pmatrix} x_1 \\ x_2 \\ ... \\ x_n \end{pmatrix}$, $\mu_i \in E(x_i)$, $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ ... \\ \mu_n \end{pmatrix}$

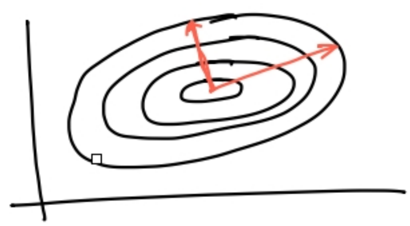
$\Sigma \in |\mathbb{R}^{n \cdot n}$ with $\Sigma_{ij} = cov(x_i, x_j)$ called the covariance matrix.

$f_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{\frac{n}{2}} |def\Sigma|^{\frac{1}{2}}}^{exp(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu))}$

Notation: $N(\mu, \Sigma)$

Prop: Σ is semi-definite and symmetric.

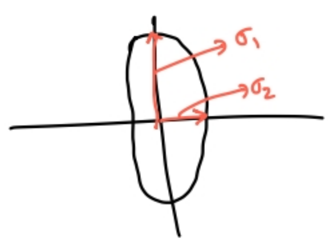
Consequence: Σ has real-valued, non-negative eigenvalues.



Contour lines of $f_{\mu,\Sigma}$

directions of eigenvectors

$X_1, X_2, ...X_n$ are independent $\Leftrightarrow \Sigma = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & ... & 0 \\ 0 & 0 & 0 & \sigma_n^2 \end{bmatrix}$



$x \sim N(\mu_1, \Sigma_1), y \sim N(\mu_2, \Sigma_2)$ independent then $x + y \sim N(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2)$

6 Mixture of Gaussians

Consider $\pi_1, \pi_2, ..., \pi_n$ with $0 \leq \pi_i \leq 1, \Sigma\pi_i = 1$

Consider the following density:

$$f(x) = \sum_{i=1}^{k} \pi_i f_{\mu_i,\Sigma_i}(x)$$