CSC2412: Algorithms for Private Data Analysis Fall 2020

Some Useful Probability Facts

Aleksandar Nikolov

Here we collect some useful probability facts. In addition to them, you should review basic probability theory, e.g., definition of a probability space, independence, conditional probability, expectation, conditional expectation.

1 Some Continuous Probability Distributions

Remember that we can define a continuous probability distribution on $\mathbb{R}^k$ (i.e., k-dimensional space) by a *probability density function* (pdf) $p : \mathbb{R}^k \to \mathbb{R}$, where p must satisfy $\forall z : p(z) \geq 0$, and

$$\int_{\mathbb{R}^k} p(z)dz = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(z)dz_1 \dots dz_k = 1.$$

The pdf p defines the probability distribution of a random variable $Z \in \mathbb{R}^k$ by

$$\mathbb{P}(Z \in S) = \int_S p(z)dz = \int_{\mathbb{R}^k} 1_S(z)p(z)dz,$$

for any (measurable) $S \subseteq \mathbb{R}^k$, where $1_S(z)$ is the function which takes value 1 on $z \in S$ and 0 on $z \notin S$.

We will use the following continuous probability distributions often:

- The *Laplace distribution* $\mathrm{Lap}(\mu, b)$ on $\mathbb{R}$ with expectation μ and scale parameter $b > 0$ has pdf

$$p(z) = \frac{1}{2b} e^{-|z-\mu|/b}.$$

- The *multivariate Laplace distribution* on $\mathbb{R}^k$ with mean $\mu \in \mathbb{R}^k$ and scale parameter $b \in \mathbb{R}, b > 0$ has pdf

$$p(z) = \frac{1}{(2b)^k} e^{-\|z-\mu\|_1/b},$$

  where $\|z - \mu\|_1 = \sum_{i=1}^k |z_i - \mu_i|$ is the ℓ_1^k norm.

- The *Gaussian (normal) distribution* $\mathrm{N}(\mu, \sigma^2)$ on R with expectation μ and variance $\sigma > 0$ has pdf

$$p(z) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(z-\mu)^2/(2\sigma^2)}.$$

- The *multivariate Gaussian distribution* $\mathrm{N}(\mu, \Sigma)$ on $\mathbb{R}^k$ with mean $\mu \in \mathbb{R}^k$ and $k \times k$ (non-singular) covariance matrix Σ has pdf

$$p(z) = \frac{1}{(2\pi)^{k/2}\sqrt{\det(\Sigma)}} e^{-(z-\mu)^\top \Sigma^{-1}(z-\mu)/2}.$$

  In particular, the spherical Gaussian is the special case when the coordinates are independent, i.e., $\Sigma = \sigma^2 I$, where I is the $k \times k$ identity matrix, and $\sigma > 0$. Then we have

$$p(z) = \frac{1}{(2\pi)^{k/2}\sigma^k} e^{-\|z-\mu\|_2^2/(2\sigma^2)},$$

  where $\|z - \mu\|_2 = \sqrt{\sum_{i=1}^k (z_i - \mu_i)^2}$ is the ℓ_2^k norm.

2 Concentration Bounds

We will also need ways to argue that a random variable is not far from its expectation. First, some bounds for Laplace and Gaussian distributions.

- If $Z \in \mathbb{R}$ is a Laplace random variable from $\mathrm{Lap}(\mu, b)$, then
$$\mathbb{P}(|Z - \mu| \geq t) = e^{-t/b}.$$
- If $Z \in \mathbb{R}$ is a Gaussian random variable from $\mathrm{N}(\mu, \sigma^2)$, then
$$\mathbb{P}(|Z - \mu| \geq t) \leq 2e^{-t^2/(2\sigma^2)}.$$
Better bounds are known, but this one is easy to prove and suffices for our purposes.

Next, some bounds that hold more generally.

- *Markov's inequality*: For any real random variable $Z \geq 0$, we have
$$\mathbb{P}(Z \geq t) \leq \frac{\mathbb{E}[Z]}{t}.$$
- *Chebyshev's inequality*: For any real random variable with expectation μ and variance σ^2, we have
$$\mathbb{P}(|Z - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$
- *Hoeffding's inequality*: for any independent random variables $Z_1, \ldots, Z_n$, such that $Z_i \in [\ell_i, u_i]$, we have
$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Z_i]\right| \geq t\right) \leq 2e^{-2n^2t^2/(\sum_{i=1}^{n}(u_i - \ell_i)^2)}.$$
In particular, if each Z_i is in $[\ell, u]$, then
$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Z_i]\right| \geq t\right) \leq 2e^{-2nt^2/(u-\ell)^2}.$$

3 Sums of Gaussians

The Gaussian distribution has many special properties. We will use one of them frequently.

- *Sum of Gaussians is Gaussian*: if $Z_1, \ldots, Z_k \in \mathbb{R}$ are jointly Gaussian, i.e., $Z = (Z_1, \ldots, Z_k)$ is distributed according to $\mathrm{N}(\mu, \Sigma)$ for some μ and Σ, then for any fixed $a_1, \ldots, a_k$, the random variable $\sum_{i=1}^{k} a_i Z_i$ is Gaussian with mean $\sum_{i=1}^{k} a_i \mu_i$ and variance $\sum_{i=1}^{k}\sum_{j=1}^{k} a_i a_j \Sigma_{i,j}$.
- As a special case, let's say $Z_1, \ldots, Z_k \in \mathbb{R}$ are independent Gaussian random variables, each with mean μ and variance σ^2. Then, for any fixed $a_1, \ldots, a_k$, the random variable $\sum_{i=1}^{k} a_i Z_i$ is Gaussian with mean $\mu \sum_{i=1}^{k} a_i$ and variance $\sigma^2 \sum_{i=1}^{k} a_i^2$.