Statistical Machine Learning II
Spring 2017, Learning Theory, Lecture 1

Jean Honorio jhonorio@purdue.edu

1 End Goal of Learning Theory

The main goal is to analyze learning problems and show that if we have enough data samples, then our experimental results for the specific data set that we used are "close" to what we would have with an infinite amount of data, with some high probability ($1-\delta$ for some small δ). On the other hand, we can also show that if we do not have enough data samples, our experimental results could be very "far" with a large probability (e.g., 50%). Similarly, we can also analyze for some specific algorithms, how many iterations are sufficient for being "close" to the optimal solution, and how many iterations are necessary for not being "far".

2 Markov's and Chebyshev's inequalities

Here we present the Markov's and Chebyshev's inequalities.

Definition 1.1. *Assume that x is a random variable with support on $\mathcal{X}$, i.e., $x \in \mathcal{X}$. p is a probability density function if and only if:*

$$(\forall x \in \mathcal{X})\ p(x) \geq 0 \quad \text{and} \quad \int_{x \in \mathcal{X}} p(x) dx = 1$$

One example of the above is the standard Gaussian random variable (with zero mean and unit variance). In that case, $\mathcal{X} = \mathbb{R}$ and $p(x) = e^{-x^2/2}/\sqrt{2\pi}$. Another example is the uniform distribution in $[0,1]$. In that case, $\mathcal{X} = [0,1]$ and $p(x) = 1$.

Definition 1.2. *The expected value of a function f of a random variable x with support on $\mathcal{X}$ and probability density function p, is defined as:*

$$\mathbb{E}[f(x)] = \int_{x \in \mathcal{X}} f(x) p(x) dx$$

With respect to the above definition, the variance of a random variable x is defined for $f(x) = (x - \mathbb{E}[x])^2$. That is, $\mathbb{V}\text{ar}[x] = \mathbb{E}[(x - \mathbb{E}[x])^2]$. The probability of an event $\mathcal{A}$ that depends on a random variable x is defined for $f(x) = 1[\mathcal{A}(x)]$. That is, $\mathbb{P}[\mathcal{A}(x)] = \mathbb{E}[1[\mathcal{A}(x)]]$. Consider for instance $\mathcal{A}(x) : x \geq a$.

Theorem 1.1 (Markov's inequality)**.** *Assume that x is a non-negative scalar random variable, i.e., $x \geq 0$. Fix $a > 0$. We have that:*

$$\mathbb{P}[x \geq a] \leq \mathbb{E}[x]/a$$

Proof.

$$\begin{aligned}
\mathbb{E}[x] &= \int_0^{+\infty} xp(x)dx \\
&= \int_0^{a} xp(x)dx + \int_a^{+\infty} xp(x)dx \\
&\geq \int_a^{+\infty} xp(x)dx \\
&\geq \int_a^{+\infty} ap(x)dx \\
&= a\int_a^{+\infty} p(x)dx \\
&= a\ \mathbb{P}[x \geq a]
\end{aligned}$$

The two inequalities follow since the probability density function fulfills the condition $(\forall x \geq 0)\ p(x) \geq 0$. □

Corollary 1.1 (Chebyshev's inequality)**.** *Assume that x is a scalar random variable, i.e., $x \in \mathbb{R}$. Fix $a > 0$. We have that:*

$$\mathbb{P}[|x - \mathbb{E}[x]| \geq a] \leq \mathbb{V}\mathrm{ar}[x]/a^2$$

Proof.

$$\mathbb{P}[|x - \mathbb{E}[x]| \geq a] = \mathbb{P}[(x - \mathbb{E}[x])^2 \geq a^2]$$

By defining the random variable $y = (x - \mathbb{E}[x])^2$ and invoking Markov's inequality (Theorem 1.1), we prove our claim. □

3 Application: Concentration Inequalities for the Mean of Gaussian variables

Next we provide concentration results for the empirical mean of standard Gaussian random variables.

Definition 1.3. *Assume that x and y are two random variables with probability density functions p_x and p_y respectively. x and y are independent if and only if their joint probability density function p_{xy} decomposes as the product of two probability density functions:*

$$(\forall x \in \mathcal{X}, y \in \mathcal{Y})\ p_{xy}(x, y) = p_x(x)p_y(y)$$

Theorem 1.2. *Assume that x and y are two independent random variables with support on $\mathcal{X}$ and $\mathcal{Y}$ respectively. Assume that $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$ are two arbitrary functions. We have that:*

$$\mathbb{E}[f(x)g(y)] = \mathbb{E}[f(x)]\ \mathbb{E}[g(y)]$$

Proof.

$$\begin{aligned}
\mathbb{E}[f(x)g(y)] &= \int_{x\in\mathcal{X}, y\in\mathcal{Y}} f(x)g(y)p_{xy}(x,y)dxdy \\
&= \int_{x\in\mathcal{X}, y\in\mathcal{Y}} f(x)g(y)p_x(x)p_y(y)dxdy \\
&= \int_{x\in\mathcal{X}} f(x)p_x(x)dx \int_{y\in\mathcal{Y}} g(y)p_y(y)dy \\
&= \mathbb{E}[f(x)]\ \mathbb{E}[g(y)]
\end{aligned}$$

□

The above definition and theorem can be easily generalized for n independent random variables.

Corollary 1.2. *Assume that $x_1 \dots x_n$ are n independent standard Gaussian random variables (with zero mean and unit variance). Fix $\varepsilon > 0$. We have that:*

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} x_i \geq \varepsilon\right] \leq e^{-n\varepsilon^2/2}$$

Proof. Pick some arbitrary $t > 0$. Note that the exponential function is non-negative. By invoking Theorem 1.1 and by independence, we have:

$$\begin{aligned}
\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} x_i \geq \varepsilon\right] &= \mathbb{P}[\sum_{i=1}^{n} x_i \geq n\varepsilon] \\
&= \mathbb{P}[e^{t\sum_{i=1}^{n} x_i} \geq e^{tn\varepsilon}] \\
&\leq \mathbb{E}[e^{t\sum_{i=1}^{n} x_i}]/e^{tn\varepsilon} \\
&= \mathbb{E}\left[\prod_{i=1}^{n} e^{tx_i}\right]/e^{tn\varepsilon} \\
&= \left(\prod_{i=1}^{n} \mathbb{E}[e^{tx_i}]\right)/e^{tn\varepsilon} \\
&= \left(\mathbb{E}[e^{tx}]\right)^n/e^{tn\varepsilon} \\
&= \left(\int_{-\infty}^{+\infty} e^{tx}e^{-x^2/2}/\sqrt{2\pi}dx\right)^n/e^{tn\varepsilon} \\
&= \left(e^{t^2/2}\right)^n/e^{tn\varepsilon} \\
&= e^{nt^2/2-tn\varepsilon}
\end{aligned}$$

In order to minimize the quadratic function $f(t) = nt^2/2 - tn\varepsilon$, we make the derivative equal to zero and solve for t. That is:

$$\begin{aligned} 0 &= \partial f(t)/\partial t \\ &= nt - n\varepsilon \end{aligned}$$

Thus, $t = \varepsilon$. Plugging this back in the above, we prove our claim. □

In fact, for any arbitrary random variable x, the *moment generating function* is defined as $\mathbb{E}[e^{tx}]$ for all $t \in \mathbb{R}$.

Theorem 1.3 (Union bound)**.** *Assume that $\mathcal{A}$ and $\mathcal{B}$ are two events that depend on a random variable x.*

$$\mathbb{P}[\mathcal{A}(x) \text{ or } \mathcal{B}(x)] \leq \mathbb{P}[\mathcal{A}(x)] + \mathbb{P}[\mathcal{B}(x)]$$

Proof. Assume that x has support on $\mathcal{X}$. We have:

$$\begin{aligned} \mathbb{P}[\mathcal{A}(x) \text{ or } \mathcal{B}(x)] &= \mathbb{E}[1[\mathcal{A}(x) \text{ or } \mathcal{B}(x)]] \\ &= \int_{x \in \mathcal{X}} 1[\mathcal{A}(x) \text{ or } \mathcal{B}(x)] p(x) dx \\ &\leq \int_{x \in \mathcal{X}} (1[\mathcal{A}(x)] + 1[\mathcal{B}(x)]) p(x) dx \\ &= \mathbb{P}[\mathcal{A}(x)] + \mathbb{P}[\mathcal{B}(x)] \end{aligned}$$

□

Next, we use Corollary 1.2 to state a two-sided bound. We also make use of a probability of error δ, which allows obtaining results in a perhaps more intuitive format.

Corollary 1.3. *Assume that $x_1 \dots x_n$ are n independent standard Gaussian random variables (with zero mean and unit variance). Fix $\delta \in (0, 1)$. We have that:*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} x_i\right| < \sqrt{\frac{2\log(2/\delta)}{n}}\right] \geq 1 - \delta$$

Proof.

$$\begin{aligned} \mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} x_i\right| < \varepsilon\right] &= \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} x_i > -\varepsilon \text{ and } \frac{1}{n}\sum_{i=1}^{n} x_i < \varepsilon\right] \\ &= 1 - \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} x_i \leq -\varepsilon \text{ or } \frac{1}{n}\sum_{i=1}^{n} x_i \geq \varepsilon\right] \end{aligned}$$

Note that if x is a standard Gaussian random variable, so is $-x$. By the union bound and Corollary 1.2, we have:

$$\begin{aligned}\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} x_i \leq -\varepsilon \text{ or } \frac{1}{n}\sum_{i=1}^{n} x_i \geq \varepsilon\right] &\leq \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} x_i \leq -\varepsilon\right] + \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} x_i \geq \varepsilon\right] \\ &= \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} -x_i \geq \varepsilon\right] + \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} x_i \geq \varepsilon\right] \\ &\leq 2e^{-n\varepsilon^2/2}\end{aligned}$$

By setting $\delta = 2e^{-n\varepsilon^2/2}$ and solving for ε, we obtain $\varepsilon = \sqrt{2\log(2/\delta)/n}$. □

4 Exercise

Prove the following (you can invoke Corollary 1.3 after a suitable change of variables):

Corollary 1.4. *Assume that $x_1 \ldots x_n$ are n independent Gaussian random variables with mean μ and variance σ^2. Fix $\delta \in (0,1)$. We have that:*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} x_i - \mu\right| < \sigma\sqrt{\frac{2\log(2/\delta)}{n}}\right] \geq 1 - \delta$$

5 Some Code for Fulfilling your Curiosity

Check if Corollary 1.3 holds experimentally. Fix $\delta = 0.01$. Draw $n = 30$ standard Gaussian random values $x_1 \ldots x_n$. Check whether $\left|\frac{1}{n}\sum_{i=1}^{n} x_i\right| < \sqrt{\frac{2\log(2/\delta)}{n}}$ or not. Repeat the above $10,000$ times.

5.1 Matlab code

```
clear all
rand('state',mod(floor(24*60*60*100*now),2^32));

delta = 0.01;
n = 100;
T = 10000;

errors = 0;
for t = 1:T
    x = randn(1,n);
    if abs(mean(x)) >= sqrt(2*log(2/delta)/n)
        errors = errors+1;
    end
```

```
end
fprintf('Probability of error (defined): %0.4f\n',delta);
fprintf('Probability of error (observed): %0.4f\n',errors/T);
```

5.2 C++ code

```
#include <cmath>
#include <cstdio>
#include <iostream>
#include <random>

int main()
{
    unsigned seed = std::chrono::system_clock::now().time_since_epoch().count();
    std::default_random_engine generator (seed);
    std::normal_distribution<double> distribution(0.0,1.0);

    double delta = 0.01;
    int n = 100;
    int T = 10000;

    int errors = 0;
    for (int t = 0; t < T; t++) {
        double s = 0;
        for (int i = 0; i < n; i++)
            s += distribution(generator);
        double mean = s/n;
        if (fabs(mean) >= sqrt(2.0*log(2.0/delta)/n))
            errors++;
    }
    printf("Probability of error (defined): %0.4f\n",delta);
    printf("Probability of error (observed): %0.4f\n",((double) errors)/T);
    return 0;
}
```