Introduction to Machine Learning

Recitation 1

Lecturer: Regev Schweiger *Scribe: Regev Schweiger*

1.1 MAP of two Gaussians

Recall that the normal, or Gaussian, distribution density function is defined as:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where μ is the expectation and σ^2 is the variance.

Suppose we are given a point $x \in \mathbb{R}$, and we need to assign a binary label to it (1 or 0). As discussed in the lecture, we assume that the data is sampled from a distribution $D(x, y)$, where x is our data point and y is its label. In addition, suppose we know the distributions $\Pr(x|y = 0)$ and $\Pr(x|y = 1)$, denoted by $f_0(x)$ and $f_1(x)$, respectively. In particular, we know that they are both Gaussian, with the same σ^2, but with different means, μ_0 and μ_1, for 0 and 1, respectively. Finally, we assume that the probability to sample a positive sample (i.e. $y = 1$) is known and we denote it by p.

Given x, what is our best guess for y? First, we need to define a suitable loss function. However, since we force y to be 0 or 1, most reasonable loss function are identical to the simple 0-1 loss function, which we will assume. We therefore would like to decide $y = 1$ if $\Pr(y = 1|x) > \Pr(y = 0|x)$. It remains to derive a simple decision condition for x. First, we use the Bayes rule for $\Pr(y = 1|x)$:

$$\Pr(y = 1|x) = \frac{\Pr(x|y = 1) \cdot \Pr(y = 1)}{\Pr(x)}$$

similarly,

$$\Pr(y = 0|x) = \frac{\Pr(x|y = 0) \cdot \Pr(y = 0)}{\Pr(x)}$$

We need not develop the denominator here, since it is cancelled:

$$\begin{aligned}
\Pr(y=1|x) > \Pr(y=0|x) &\Leftrightarrow \frac{\Pr(x|y=1)\cdot\Pr(y=1)}{\Pr(x)} > \frac{\Pr(x|y=0)\cdot\Pr(y=0)}{\Pr(x)} \\
&\Leftrightarrow \frac{\Pr(x|y=1)}{\Pr(x|y=0)} > \frac{\Pr(y=0)}{\Pr(y=1)} \\
&\Leftrightarrow \frac{f_1(x)}{f_0(x)} > \frac{1-p}{p} \\
&\Leftrightarrow \frac{\frac{1}{\sqrt{2\pi\sigma^2}}\cdot e^{-\frac{(x-\mu_1)^2}{2\sigma^2}}}{\frac{1}{\sqrt{2\pi\sigma^2}}\cdot e^{-\frac{(x-\mu_0)^2}{2\sigma^2}}} > \frac{1-p}{p} \\
&\Leftrightarrow e^{-\frac{(x-\mu_1)^2-(x-\mu_0)^2}{2\sigma^2}} > \frac{1-p}{p} \\
&\Leftrightarrow -\frac{(x-\mu_1)^2-(x-\mu_0)^2}{2\sigma^2} > \log\left(\frac{1-p}{p}\right) \\
&\Leftrightarrow (x-\mu_1)^2-(x-\mu_0)^2 < -2\sigma^2\log\left(\frac{1-p}{p}\right) \\
&\Leftrightarrow (x^2-2x\mu_1+\mu_1^2)-(x^2-2x\mu_0+\mu_0^2) < -2\sigma^2\log\left(\frac{1-p}{p}\right) \\
&\Leftrightarrow -2x(\mu_1-\mu_0)+(\mu_1^2-\mu_0^2) < -2\sigma^2\log\left(\frac{1-p}{p}\right) \\
&\Leftrightarrow -2x(\mu_1-\mu_0) < -2\sigma^2\log\left(\frac{1-p}{p}\right)-(\mu_1^2-\mu_0^2) \\
&\Leftrightarrow x > \log\left(\frac{1-p}{p}\right)\cdot\frac{\sigma^2}{\mu_1-\mu_0}+\frac{\mu_1+\mu_0}{2}
\end{aligned}$$

In summary, we have seen that there exists a constant C, such that if $x > C$, we will decide that $y = 1$. Further inspection shows us that as p (the probability of sampling $y = 1$) is higher, the lower of a threshold we need to decide $y = 1$, and vice versa.

1.2 Concentration Bounds

In probability theory, concentration inequalities provide bounds on how a random variable deviates from some value (typically, its expected value). As we go through the several bounds we have learned, we will apply each of them to the case of biased coins. Let $X_1, \ldots, X_N$ be i.i.d random variables, where $X_i \sim \mathit{Bernoulli}(p)$ - n flipped coins, each with probability

p to get ”heads”. What can we say about the number of times we got ”heads”? Or, to make it easier on ourselves, we instead look at the average: $X = \frac{1}{n}\sum_{i=1}^{n} X_i$. Its expectation is $\mathrm{E}(X) = p$; what can be say about the probability to be far from the expectation? In particular, we are interested in:

$$\Pr(X \geq p + \varepsilon), \quad \Pr(X \leq p - \varepsilon)$$

1.2.1 Markov inequality

We will make different assumptions on X, and get gradually better bounds. First, we will only assume $X \geq 0$. The Markov inequality is

$$\Pr(X \geq a) \leq \frac{\mathrm{E}(X)}{a}$$

The proof is simple; expand the expectation to get:

$$\begin{aligned}\mathrm{E}(X) &= \sum_{x=0}^{\infty} x \cdot \Pr(X = x) \\ &\geq \sum_{x=a}^{\infty} x \cdot \Pr(X = x) \\ &\geq a \cdot \sum_{x=a}^{\infty} \Pr(X = x) = a \cdot \Pr(X \geq a)\end{aligned}$$

Application to biased coin flips. Here, $\mathrm{E}(X) = p$, and $a = p + \varepsilon$ (we can’t get a bound on the other side here easily). We get:

$$\Pr(X \geq p + \varepsilon) \leq \frac{p}{p + \varepsilon} = O\left(\frac{1}{\varepsilon}\right)$$

which is a linear tail bound.

1.2.2 Chebyshev inequality

Now, we drop the positivity assumption, but instead assume that the variance is finite and known - σ^2, the Chebyshev inequality is:

$$\Pr(|X - \mathrm{E}(X)| \geq b) \leq \frac{\sigma^2}{b^2}$$

The proof is simple: Define $Y = (X - \mathrm{E}(X))^2$ and $a = b^2$, and apply the Markov inequality to get:

$$\begin{aligned}\Pr(Y \geq a) &\leq \frac{\mathrm{E}(Y)}{a} \Rightarrow \\ \Pr((X - \mathrm{E}(X))^2 \geq a) &\leq \frac{\mathrm{E}((X - \mathrm{E}(X))^2)}{a} \Rightarrow \\ \Pr(|X - \mathrm{E}(X)| \geq b) &\leq \frac{\sigma^2}{b^2} \Rightarrow\end{aligned}$$

Application to biased coin flips. Here, $\mathrm{Var}(X) = p(1-p)/n$. For the two-tailed inequality, we can get:

$$\Pr(X \geq p + \varepsilon) + \Pr(X \leq p - \varepsilon) = \Pr(|X - p| \geq \varepsilon) \leq \frac{p(1-p)}{n\varepsilon^2} = O\left(\frac{1}{\varepsilon^2}\right)$$

which is a quadratic bound, but still polynomial.

1.2.3 Hoeffding inequality

Now we assume that we have a sum of i.i.d. Bernoulli variables, $S = X_1 + \ldots + X_n$, where X_i are defined as before, and $\mathrm{E}(S) = np$. The Hoeffding inequality is:

$$\begin{aligned}\Pr(S \geq \mathrm{E}(S) + c) &\leq e^{-\frac{2c^2}{n}} \\ \Pr(S \leq \mathrm{E}(S) - c) &\leq e^{-\frac{2c^2}{n}}\end{aligned}$$

An outline of the proof is available in the lesson scribes. This is a slightly weaker bound than the more general Chernoff bound.

Application to biased coin flips. Here, $X = S/n$. So, we can get:

$$\begin{aligned}&\Pr(X \geq p + \varepsilon) = \Pr(S/n \geq p + \varepsilon) = \Pr(S \geq np + n\varepsilon) \leq e^{-\frac{2(n\varepsilon)^2}{n}} = e^{-2\varepsilon^2 n} \\ &\Pr(X \leq p - \varepsilon) \leq e^{-2\varepsilon^2 n}\end{aligned}$$

which is an exponential bound.

1.2.4 Chernoff bounds

We cite two more bounds. The additive form (absolute error) Chernoff bound: Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli variables, $X_i \sim Bernoulli(p)$, and $X = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then:

$$\Pr\left(X \geq p+\varepsilon\right) \leq \left(\left(\frac{p}{p+\varepsilon}\right)^{p+\varepsilon}\left(\frac{1-p}{1-p-\varepsilon}\right)^{1-p-\varepsilon}\right)^{n}$$

$$\Pr\left(X \leq p-\varepsilon\right) \leq \left(\left(\frac{p}{p-\varepsilon}\right)^{p-\varepsilon}\left(\frac{1-p}{1-p+\varepsilon}\right)^{1-p+\varepsilon}\right)^{n}$$

The multiplicative form (relative error) Chernoff bound: Let $S = X_1 + \ldots + X_n$ be the sum of n independent – but not necessarily identically distributed – variables, and $\mu = \mathrm{E}(S)$. Then:

$$\Pr(S \geq (1+\delta)\mu) \leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}$$

$$\Pr(S \leq (1-\delta)\mu) \leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu}$$

1.2.5 A numerical example

Suppose we flip 100 unbiased coins. Can we bound the probability that more than 3/4 of them will fall on "heads"? Formally, we have $p = 0.5, \varepsilon = 0.25, n = 100$. Then:

1. The Markov inequality gives us $\Pr(X \geq p+\varepsilon) \leq 0.5/0.75 = 2/3$.

2. The Chebyshev inequality gives us $\Pr(X \geq p+\varepsilon) \leq \Pr(X \geq p+\varepsilon) + \Pr(X \leq p-\varepsilon) = \Pr(|X-p| \geq \varepsilon) 0.5^2/(100 \cdot 0.25^2) = 0.04$.

3. The Hoeffding inequality gives us $\Pr(X \geq p+\varepsilon) \leq e^{-2 \cdot 0.25^2 \cdot 100} \approx 0.0000037$.