Math Rules for Machine Learning

Aimee Barciauskas

29 March 2016

Misc

- $min(a, b) = \frac{a+b-|a-b|}{2}$
- $1 + x \leq e^x$
- $e^{\lambda x} \leq xe^{\lambda} + (1 - x)$

Probabilities

- $\mathbb{P}\big\{Bin(n, p) = k\big\} = \binom{n}{k} p^k (1 - p)^{n-k}$
- **Union of Events Bound**: $\mathbb{P}\big(\bigcup_j A_j\big) \leq \sum_j \mathbb{P}(A_j)$

Expected Values

- $\mathbb{E}[X] = \int x f(x) dx,\ X$ admits a pdf $f(x)$
- **Linearity of expectation**: $\mathbb{E}(X) = \frac{1}{n} \sum_i^n X_{i,n-1}$
  - Linearity of expectation is the property that the expected value of the sum of random variables is equal to the sum of their individual expected values, regardless of if they are independent.
- **Identically Distributed**: expected value of a random variable is equal to the average of identically distributed random draws: $\frac{1}{n} \sum X_i = \mathbb{E}X$
- $Var(X) = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big]$

Inequalities

- **Markov**: $\mathbb{P}\big\{X \geq t\big\} \leq \frac{\mathbb{E}[X]}{t}$
- **Chebyshev**: $\mathbb{P}\big\{\ |X - \mathbb{E}[X]|\ \big\} \leq \frac{var(X)}{t^2}$

  - Reveals that typical deviations from the expected value are of the order $\frac{\sigma}{\sqrt{(n)}}$
  - $\mathbb{P}\{\,|X - \mathbb{E}[X]| \geq k\sigma\} \leq \frac{1}{k^2}$
- **Chernoff**: $\mathbb{P}\{X \geq a\} \leq \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda a}}$
  - When X is a sum of n random variables the RHS $= \frac{\mathbb{E}\left[\prod_{i=1}^{n} e^{\lambda X_i}\right]}{e^{\lambda a}}$
- **Jensen's Inequality**: $\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)]$
  - Used when bounding the expected value of an empirical frequency from it's expectation (Lecture 4, slide 2)
- *Problem 16 Solution*: $\mathbb{P}\left\{X - \mathbb{E}X \geq nx\right\} \geq e^{-nx^2}$
- **Cauchy-Schwarz**: $|\langle x, y\rangle| \leq \|x\|\dot{\|}y\|$

VC dimensions and shatter coefficients

*In general, to prove the VC dimension of a class $\mathcal{A}$ it is sufficient to show k is the VC dimension if $k+1$ points can be shattered

- $\mathcal{A}$ is a class of sets with VC dimension $V_{\mathcal{A}}$, then for every n: $s(\mathcal{A}, n) \leq \sum_{i=1}^{V_{\mathcal{A}}} \binom{n}{i}$ (Theorem 13.2)
- For all $n > 2V$, $s(\mathcal{A}, n) \leq \left(\frac{en}{V_{\mathcal{A}}}\right)^{V_{\mathcal{A}}}$
- If $\mathcal{A}$ contains finitely many sets, then $V_{\mathcal{A}} \leq log_2 |\mathcal{A}|$ and $s(\mathcal{A}, n) \leq |\mathcal{A}|$ for every n (Theorem 13.6)
  - Proof: The first inequality follows from the fact that at least 2^n sets are necessary to shatter n points. The second inequality is trivial.
- $\mathcal{A}$ is set of all half-lines $(-\infty, x]$: $s((A), 2) = 3 < 2$, so $V_{\mathcal{A}} = 1$ and $s(\mathcal{A}, n) = n + 1 = \binom{n}{0} + \binom{n}{1}$
  - Proof: Any 2 different points $z_1 < z_2$, there is no set of the form that contains z_2 and not z_1
- $\mathcal{A}$ is set of all half-intervals, $V_{\mathcal{A}} = 2$ and $s(\mathcal{A}, n) = \frac{n(n+1)}{2} + 1$
  - Proof: To see that the vc dimension is 2, observe that if we fix 3 different points in $\mathcal{R}$, then there is no interval that does not contain the middle point but does contain the other 2. The shatter coefficient can be calculated by counting that there are at most $n - k + 1$ sets in A intersection of x1,. . . ,x_{n} such that the absolutve number is k, for $k = 1, ..., n$and one set where this is 0.

- In $\mathcal{R}^d$:
  - half-lines: $V_{\mathcal{A}} = d$
  - all rectangles: $V_{\mathcal{A}} = 2d$
- $\mathcal{A}$ set of halfspaces in $\mathcal{R}^d$ of the form $\{x : ax \geq b\}$, $V_{\mathcal{A}} = d + 1$ and $s(\mathcal{A}, n) = 2\sum_{i=0}^{d} \binom{n-1}{i} \leq 2(n-1)^d + 2$ (Corollary 13.1)
  - Proof: If we take G to be the linear space spanned by d functions $x^{(d)}$ and the $d + 1$ function $= 1$, where $x^{(d)}$ is the d-th component of x

Theorem 13.9

Let $\mathcal{G}$ be a finite-dimensional vector space of real functions on $\mathcal{R}^d$. The class of sets:

$$\mathcal{A} = x : g(x) \geq 0 : g \in G$$

*has VC dimension $V_{\mathcal{A}} \leq r$, where r is the dimension of G

The class of sets of this form have:

$$s(\mathcal{A}, n) \leq \sum_{i=0}^{r} \binom{n}{i}$$

In many cases it is possible to get sharper estimates, let $\mathcal{G}$ be the linear space of functions spanned by some fixed functions $\psi_1, ..., \psi_r$, if every r-element subset is linearly independent, then the n-th shatter coefficient is $\mathcal{A} = \{\{x : g(x) \geq 0\} : g \in \mathcal{G}\}$ actually equals:

$$s(\mathcal{A}, n) = 2\sum_{i=0}^{r-1} \binom{n-1}{i}$$