Math Rules for Machine Learning

Aimee Barciauskas

29 March 2016

Misc

- $min(a,b) = \frac{a+b-|a-b|}{2}$
- $min(a,b) \leq \sqrt{ab}$
- $1 + x \leq e^x$
- $e^{\lambda x} \leq xe^{\lambda} + (1-x)$
- The number of permuations of k objects from a set n: $\frac{n!}{(n-k)!}$
- **Probably Approximately Correct**: $\mathbb{P}\{R(g_n) - R^* < \epsilon\} \geq 1 - \delta$

Probabilities

- $\mathbb{P}\{X \in A\} = \mathbb{E}[\mathbb{I}_{X \in A}]$
- $\mathbb{P}\big\{Bin(n,p) = k\big\} = \binom{n}{k} p^k (1-p)^{n-k}$
- **Union of Events Bound**: $\mathbb{P}\big(\bigcup_j A_j\big) \leq \sum_j \mathbb{P}(A_j)$

Expected Values

- $\mathbb{E}\mathbb{E}[X|Y] = X$ i.e. a value
- $\mathbb{E}[X] = \int xf(x)dx,\ X$ 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: $\frac{1}{n}\sum X_i = \mathbb{E}X$
- $Var(X) = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big]$
- Independence: $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$

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 $\frac{\sigma}{\sqrt(n)}$
  - $\mathbb{P}\big\{\ |X - \mathbb{E}[X]| \geq k\sigma\big\} \leq \frac{1}{k^2}$
- **Chernoff**: $\mathbb{P}\big\{X \geq a\big\} \leq \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda a}}$
  - When X is a sum of n random variables the RHS $= \frac{\mathbb{E}\Big[\prod_{i=1}^n e^{\lambda X_i}\Big]}{e^{\lambda a}}$
- **Jensen's Inequality**: $\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)]$

  - e.g. $e^{\mathbb{E}[X]} \leq \mathbb{E}e^x$
  - Used when bounding the expected value of an empirical frequency from it's expectation (Lecture 4, slide 2)

- *Problem 16 Solution*: $\mathbb{P}\Big\{X - \mathbb{E}X \geq nx\Big\} \geq e^{-nx^2}$
- **Cauchy-Schwarz**: $|\langle x, y\rangle| \leq \|x\|\|y\|$
- **Sauer's Lemma**: For all n: $s(\mathcal{A}, n) \leq (n+1)^V$
- By convexity of $e^{\lambda x} \leq xe^{\lambda} + (1-x)$

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: $V_{\mathcal{A}} = 2d$
- Let $\mathcal{A}$ be the 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
- Linear classifiers: $d + 1$

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:

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

- The class of all convex polygons has infinite VC dimension.
- The class of all polygons with k vertices in the plane.
    - Solution: In order to show that the VC-dimension of a class of concepts is d, we need to show that there exists a set of size d on which all dichotomies are realized, and that on all sets of size d + 1, there is some dichotomy that is not realized by the concept class. We know from class that the class of convex polygons with k vertices has VC dimension of 2k + 1. Thus, we know that the VC dimension of our class is at least 2k + 1. It can be shown that for all sets of size 2k + 2 points, there is a labeling of these points that cannot be captured by (even) non-convex polygons with k vertices.
- The class of all circles in the plane.
    - Solution: It is clear that any two points can be shattered by a circle. Any three non-colinear points can also be shattered. Now, given any four points, there are two cases. The first, that the convex hull of these four points is a triangle. If so, labelling the points on the triangle as positive and the point inside as negative is a dichotomy that cannot be realized by a circle. If the convex hull of the four points is a quadrilateral, then choosing the further of the two diagonally opposite points as positive and the other two as negative is a dichotomy that cannot be realizeda lso. Finally, if the four points are colinear, there is a trivial dichotomy of alternate positive and negatives that cannot be realized. Thus, VC dimension of all circles in the plane is 3.
- The class of union of k intervals on the real line.
    - Solution: Easy to check that a sequence of $2k+1$ points on a line cannot be shattered, if successive points are labeled with alternate labels, starting with a positive label. Thus, VC dimension of the class of union of k intervals on the real line is 2k.
- Let F be a finite-dimensional vector space of real functions on R_n, $dim(F) = r < \infty$. Let H be the set of hypotheses: $H = x : f(x) \geq 0 : f \in F$. Show that d, the VC dimension of H, is finite and that $d \leq r$ [Hint: select an arbitrary set of $m = r + 1$ points and consider the linear mapping $u : F \rightarrow R_m$ defined by: $u(f) = (f(x_1), ..., f(x_m))$.]
- Solution: Show that no set of size $m = r + 1$ can be shattered by H. Let $x_1, ..., x_m$ be m arbitrary points. Define the linear mapping $l : F \rightarrow R_m$ defined by:

  $l(f) = (f(x_1), ..., f(x_m))$

  Since the dimesion of $dim(F) = m - 1$, the rank of l is at most $m - 1$ and there exists $\alpha \in R_m$ orthogonal to $l(F)$:

  $\forall f \in F, \sum_{i=1}^{m} \alpha_i f(x_i) = 0$

  We can assume at least one α_i is negative. Then,

$\forall f \in F, \sum_{i:\alpha_i \geq 0}^{m} \alpha_i f(x_i) = -\sum_{i:\alpha_i < 0}^{m} \alpha_i f(x_i)$

Now, assume that there exists a set $x : f(x) \geq 0$ selecting exactly the xis on the left-hand side. Then all the terms on the left-hand side are non-negative, while those on the right-hand side are negative, which cannot be. Thus, $(x_1, ..., x_m)$ cannot be shattered.