Machine Learning Problem Sets

Aimee Barciauskas

29 March 2016

Set 1

Problem 1: Determine the Bayes Risk using class-conditional probabilities

$R^* = min(\eta(x), 1 - \eta(x))$

- Use: $\eta(x) = \frac{\mathbb{P}\{Y=1\} f_1(x)}{f(x)}$, where $f_1(x)$ is the conditional distribution of x given $Y = 1$
- Substitute prior probabilities, remove normalizing $f(x)$

Problem 2: Determine the Bayes Classifier using class-conditional probabilities

$$g^*(x) = \begin{cases} 1 & \text{if } q_1 f_1 > q_0 f_0 \\ 0 & \text{otherwise} \end{cases}$$

- Substituted f_1 and f_0 directly, reduce to find when the inequality holds

Problem 3 and 4: Determine a function which is the optimal classifier for a given loss function

- The function minimizes the expected loss: $\mathbb{E}[\ell(y, y')] = \int \ell(y, y') p(y|x) dy$
  - Take derivative wrt to function y', set equal to 0 and solve for y'
- Show that for any alternative function, the expected loss is greater than or equal to the loss of the function.

Problem 5: Show the probability the distance of the k-nearest neighbors is 0 goes to 1 as n goes to infinity.

- Show the probability that the distance is greater than epsilon goes to 0:
  1. Take the expected value of the probability wrt X
  2. Replace the probability expression with it's binomial expression: the probability that n samples are not in the ball is less than k
  3. This is less than or equal to the unconditional probability
  4. Solve the probability using the expression for the probability of a binomial:

  $\mathbb{P}\big\{Bin(n, q) < k\big\} = \sum_{i=1}^{k-1} nchooseiq^k (1-q)^{n-k}$

  5. This expression goes to zero with $n \to \infty$

Problem 6: Show $\mathbb{E}[R(g)] > \frac{1}{4}$ even when $R^* = 0$

- For 1-NN: Show the probability that nothing falls in the X_i's "bucket" is greater than $\frac{1}{4}$
  - e.g. $\mathbb{P}\big\{Bin(n, \frac{1}{m}\big\} \geq 2$

Set 2

Problem 7: Calculate R^*, R_{1-NN}, R_{3-NN}

- Expand R as a function of $\eta(x)$ and take expected value (e.g. integrate over possible values of $eta(x)$)

Problem 8: Show that the probability a sum of random indpendent variables taking values in $[0, 1]$ is greater than t is bounded by a complicated function of it's expected value and t

1. Use Chernoff to rewrite in terms of expectations and a product over X_i's
2. Use convexity of $e^{\lambda x}$ to take x out of exponent
3. Bring through expected value of X_i
4. Use $1 + x \leq e^x$ to put expected value back in exponent, now can use sum of expected value in exponent (which will just be the expected value)
5. This will be some exponential function, which will be minimized when the exponent is minimized. So take derivate and set equal to lambda. Put lambda back in and reduce.

Problem 9: Show the deviation from R^* of $R(g)$

1. Restate R in terms of the expected risk, i.e.: probability of each type of error by the probability of each class $\eta(x)$, $1 - \eta(x)$
2. Is it possible to simplify to an R^* term? Now we have an expression for the deviance from R^*
3. Use Hoeffding to replace probabilty a random variable deviates from it's expected value by more than some expression, say t, is $\leq e^{-2nt^2}$

Problem 10: Prove "structural" results of some sets (e.g. Rademacher averages)

- Rewrite structural result in terms of full expression, note if it's an inquality and there is some supremum of infinum involved it could be trivial

Problem 11: Determine the n-th shatter coefficient

$s(\mathcal{A}, n) = max_{(z_1,...,z_n) \in \{\mathcal{R}^d\}} N_{\mathcal{A}}(z_1, ..., z_n)$ where N is the number of different sets which in union compose $\mathcal{A}$

The shatter coefficient is the maximal number of different subsets of n points that can be picked out by the class of sets $\mathcal{A}$

Set 3

Problem 12: What is the VC dimension of $f(x)$

- When infinite, it suffices to prove that for any n there exist a set $x_1, ..., x_n$ such that these points may be shattered by $f(x)$
  1. Define the sequence of $x_1, ..., x_n$ (e.g. 2^{-n})
  2. Show that no matter what the assignment of $(y_1, ..., y_n)$, we can define a classification function of $f(x)$ which assigns them to that set of $\{0, 1\}$

Problem 13: Upper bound the VC dimension of the union of k classes, each having VC dimension of at least V

1. By sauer's lemma, we know the shatter coefficient of each class is bounded above by $(n+1)^V$, for the union we know this is again upper-bounded by $k(n+1)^V$
2. The VC dimension must be less than 2^n, so we can reduce to an expression for n: $n > ...$
3. $n < n+1$: use the hint that with $a \geq 1$ and $b > 0$, if $x \geq 4alog(2a) + 2b$ then $x \geq alog(x) + b$
4. Subtract 1 from both sides, VC dimension cannot be greater than that value

Problem 14: Prove the perceptron algorithm converges at a rate dependent on the distance of the initial weight vector to the optimal (squared).

1. Re-express and expand $|w_t - w_*|^2$ using $w_t = w_{t-1} + \frac{Y_t X_t}{|X_t|}$
2. Simplify the expression taking note of $w_{t-1}^T X_t Y_t \leq 0$ and $w_*^T X_t Y_t \geq 1$
3. Step backward in time to notice that for all values of t: $|w_t - w_*|^2 \leq |w_{t-1} - w_*|^2 - 1$

Problem 15: Derive a bound for the expected risk of the perceptron algorithm using the leave one out estimator

1. The expected risk can be estimated by an average of the risk of n leave-one-out classifiers
2. We can use this to estimate the risk of perceptron using the rate of convergence defined in problem 14, given that we know how many times the algorithm made a mistake, which is in effect one instance of the n leave-one-out classifiers.

Problem 16: What is the expected risk of the majority classifier?

1. Re-express the expected value of the risk as the probability that the majority says 1 and the true value is 0 plus the complement.
2. Re-express the probabilities in terms of binomials in n and p

Set 4

Problem 17: Formulate a convex optimization problem for w^* and show it maximizes the margin

Actual Solution

By definition of the margin:

$$\frac{y_i w^T x_i}{\|w\|\gamma(w)} \geq 1$$

Let

$$v = \frac{w}{\|w\|\gamma(w)}$$

Maximizing w is equivalent to minimizing $\frac{1}{\gamma(w)}$, the optimization problem becomes $min\|v\|$ subject to $y_i w^T x_i \geq 1$

This is a convex optimization problem with linear constraints.

To show the optimal v lies in the subspace spanned by those x_i for which $y_i v^T x_i = 1$ (the support vectors), suppose it's not true. Then v has a component in the orthogonal complement of those x_i. By projecting orthogonally to the subspace spanned by the support vectors, the projection $\tilde{v}$ satisfies $\tilde{v}^T x_i = v^T x_i$ for all support vectors - and therefore has the same margin, but $\|\tilde{v}\| \leq \|v\|$, contradicting the optimality of v

Problem 18: Show the data are linearly seperable when Y is 1 whenever at least 1 x_i is 1.

Seperate those points that are all zero from the rest by a hyperplane by defining a classifer: $x^T \mathbf{1} \geq c$ for any $c \in (0, 1)$ where $\mathbf{1}$ is the all 1's vector.

The distance of these points to the plane is $\frac{c}{d}$. To maximize this distance, set $c = \frac{1}{2}$ and the margin is $\frac{2}{\sqrt{d}}$

Problem 19: Determine the kernel function for a feature mapping

Take the dot product of the kernel function:

$K(x, y) = \langle \phi(x), \phi(y) \rangle$

and evaluate it: re-express and simplify

The generalization to $\mathbb{R}^d$ just uses vectors and the norm of the distance: $\|\mathbf{x} - \mathbf{y}\|$

Problem 20: Show some structural manifestation of multiple kernels is also a kernel function

Show the structural result required is positive semi-definite

Additional Exercises

Problem 1

First part: When X and Y are independent, then $R^* = min(\eta(x), 1-\eta(x))$ can be reduced to the unconditional probability in $Y = 1$ *Second part:* Idea: When class conditional probabilities are the same?

Problem 2

On paper

Problem 3

See theorem 32.4

Problem 4

On paper

Problem 5

On paper but unfinished

Problem 6

See Admissibility of the Nearest Neighbor Rule

Problem 7

scale-invariant nearest neighbor rule

The scale-invariant k-nearest neighbor rule is based upon empirical distances that are defined in terms of the order statistics along the d coordinate axes. First order the points x, $X_1, ..., X_n$ according to increasing values of their first components $x^{(1)}$, $X_1^{(1)}, ..., X_n^{(1)}$, breaking ties via randomization. Denote the rank of $X_i^{(1)}$ by $r_i^{(1)}$, and the rank of $x^{(1)}$ by $r^{(1)}$. Repeating the same procedure for the other coordinates, we obtain the ranks:

$r_i^{(j)}, r^{(j)}, j = 1, ..., d, i = 1, ..., n$

Define the empirical distance between x and X_i by:

$\rho(x, X_i) = \max_{1 \leq j \leq d} \left| r_i^{(j)} - r^{(j)} \right|.$

A $k-nn$ rule can be defined based on these distances, by a majority vote among the Y_i's with the corresponding X_i's whose empirical distance from x are among the k smallest. Since these distances are integer-valued, ties frequently occur. These ties should be broken by randomization.

Problem 8

On paper

Problem 9

See Set 3, problem 16.

Problem 10

[Problem of all squares]

Problem 11

All sets in $\mathbb{R}$ as the union of k closed intervals has VC dimension $2k$

Proof:

If $k = 1$, the VC dimension is 2. Pts 1 and 3 cannot be containted in an interval not containing 2.

If $k = 2$ the VC dimension is 4 because 1, 3, and 5 cannot be contained in intervals not containing 2 or 4.

In general, $2k + 1$ points cannot be shattered by $2k$ intervals because a disjoint set of $2k + 1$ points cannot be shattered by the $2k$ intervals.

Problem 13

Set 4, Problem 17

Problem 14

Consider the cost functional $A(f) = E\phi(-f(X)Y)$ where $\phi : R \to R_+$ is a positive, increasing, strictly convex cost function, $f : X \to \mathbb{R}$ is a real-valued function and $Y \in \{-1, 1\}$. Determine the function f^* that minimizes $A(f)$. Show that the classifier $g(x) = sgn(f^*(x))$ is the Bayes classifier.

$\min_{f(X)} \mathbb{E}[\phi(-f(x)Y)]$

$\min_{f(X)} \int f(x)\phi(-f(x)Y)dx$

Problem 15

See Math Rules

Problem 16

Set 4, Problem 18