15-359: Probability and Computing

Assignment 8 Due: April 6, 2012

Problem 1: Upper tail (10 pts.)

Starting with the strongest form of the Chernoff bound,

$$\mathrm{P}(X \geq (1+\delta)\mu) < \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu},$$

prove the simplified form $\mathrm{P}(X \geq (1+\delta)\mu) < e^{-\mu\delta^2/3}$ (assuming $0 < \delta \leq 1$).

Problem 2: The $\frac{1}{2}$-th moment (10 pts.)

Suppose $X \sim \mathrm{Exp}(1)$. Compute $\mathrm{E}(\sqrt{X})$ (the $\frac{1}{2}$-th moment of X). You will need to use the integral of the Normal pdf.

(Hint: you will need to use both integration by parts and a u-substitution before getting the integral in the right form. It's a kitchen sink integral exercise.)

Problem 3: Sorting out loose ends (20 pts.)

Earlier, we had proved that randomized quicksort runs in expected $\Theta(n \log n)$ time. Using Chernoff bounds, we can go further. Let T is the running time of randomized quicksort on an input of size n. Then $\mathrm{P}(T > Cn \log n)$ drops very sharply as $n \to \infty$. By choosing an appropriate C, we can get a bound of n^{-k} on this probability, for any k.

To obtain this result, proceed as follows. Let $(a_1, \ldots, a_n)$ be the unsorted input. We track the progress of an arbitrary element $a = a_j$. Define the indicator variable X_i to equal 1 if, when we recurse on a sub-array containing a for the i-th time, the element a ends up in the smaller half.

A. Show that if $\sum_{i=1}^{t} X_i$ exceeds $\log_2 n$, then the element a must be sorted after at most t steps.

B. Use a Chernoff bound to bound the probability that $\sum_{i=1}^{t} X_i$ is less that $\log_2 n$, if $t = C \log_2 n$.

C. Find the probability that every element is sorted after $C \log_2 n$ steps (and therefore, that $T < Cn \log_2 n$). What value of C will guarantee that this probability is less than n^{-2}?

Problem 4: From real to binary (20 pts.)

Background

Suppose A is an $n \times n$ 0/1-matrix and $x \in [0,1]^n$ a vector of reals in $[0,1]$. We would like to replace x by a binary "approximation" $\widehat{x}$ such that $\|A(\widehat{x}-x)\|_\infty$ is minimal. Recall that the Chebyshev norm $\|z\|_\infty$ of a vector z is the maximum of the absolute values of its components.

One can show that this problem is NP-hard, so it is natural to look for a randomized algorithm:

1. Define indicator random variables X_i such that $\mathrm{P}(X=1) = x_i$.
2. Set $\widehat{x} = (X_1, X_2, \dots, X_n)$.

For simplicity write $Z = A(\widehat{x}-x)$.

Task

A. Warmup: Let U_i, $i = 1, \dots, n$ be independent random variables such that $\mathrm{P}(U_i = 1-p_i) = p_i$ and $\mathrm{P}(U_i = -p_i) = 1-p_i$, $0 \le p_i \le 1$. Set $U = \sum U_i$. Show that $\mathrm{P}(|U| \ge a) \le 2e^{-2a^2/n}$.
Hint: use $p_i e^{t(1-p_i)} + (1-p_i)e^{-tp_i} \le e^{t^2/8}$.

B. Determine the expectation of Z_1.

C. Bound the tail probability $\mathrm{P}(Z_1 \ge 2\sqrt{n \ln n})$.

D. Show that with probability $O(1/n)$ we have $\|Z\|_\infty < \sqrt{n \ln n}$.

Problem 5: Predictions are guesses anyway (8 pts.)

In stochastic optimization, we pick a distribution Λ over H rather than a single $h \in H$. Suppose h is drawn from Λ. Give a reasonable condition under which

$$P(F(h) - F(\Lambda) > \varepsilon)$$

can be upper-bounded by an exponential concentration inequality like Hoeffding or bounded differences.

Problem 6: PB wrap (15 pts.)

Let's complete the proof of the PAC-Bayes inequality. Fill in the blanks in the following mad-lib.

2. Let $s = F_m(h)$ and $d = F(h)$.

$$\begin{aligned}\mathrm{E}_S(Z) =& \mathrm{E}_S\left(\mathrm{E}_{h\sim\Pi}\left(\exp\left(m\left(s\ln\left(\frac{s}{d}\right) + (1-s)\ln\left(\frac{1-s}{1-d}\right)\right)\right)\right)\right)\\ =& \mathrm{E}_S\left(\mathrm{E}_{h\sim\Pi}\left(\left(\frac{s}{d}\right)^{m\cdot s}\left(\frac{1-s}{1-d}\right)^{m(1-s)}\right)\right)\end{aligned}$$

s can take ___ values: {___} and has a _________ distribution with parameter(s) ___. Thus,

$$\begin{aligned} E_S\left(\left(\frac{s}{d}\right)^{m\cdot s}\left(\frac{1-s}{1-d}\right)^{m(1-s)}\right) &= \sum_{k=0}^{m}\binom{m}{k}d^k(1-d)^{m-k}\left(\frac{k/m}{d}\right)^k\left(\frac{1-k/m}{1-d}\right)^{m-k} \\ &= \sum_{k=0}^{m}\binom{m}{k}\left(\frac{k}{m}\right)^k\left(1-\frac{k}{m}\right)^{m-k} \end{aligned}$$

___________. *Thus* $\mathrm{E}_S(Z) \le m+1$.

Problem 7: Train and test (17 pts.)

Consider a stochastic optimization problem in which $f(h) \in \{0,1\}$ and the distribution D is over a finite set of functions. We pick a Λ and then we draw a $\hat{h}$ from Λ. To bound $F(h)$, we could use question 1 in conjunction with the PAC-Bayes inequality from class. In this question, we bound $F(h)$ by h's performance on a 'test set' T of n more data drawn independently from D. That is, we calculate $F'_n(h) = \frac{1}{n}\sum_{f\in T} f(h)$.

A. Upper bound $\mathrm{P}(F(h) - F'_n(h) > \varepsilon)$.

B. We have a 'shifted' test set $\tilde{T}$ drawn from a different distribution $\tilde{D}$. Fortunately, $\tilde{D}$ is close: it's defined over the same set, and for all f, $\frac{\tilde{D}(f)}{D(f)} < r$ for some small r. We observe $\tilde{F}'_n(h) = \frac{1}{n}\sum_{f\in\tilde{T}} f(h)$. Prove an upper bound on $\mathrm{P}(F(h) - \tilde{F}'_n(h) > \varepsilon)$.

   (Hint: show expectations are close, then use a concentration inequality to show the outcome is close to the expectations. The bound may depend on $\tilde{F}(h)$.)

Problem 8: Linear classifiers (extra credit) (10 pts.)

Stochastic optimization with linear functions is an important special case. Let $\mathcal{H} = \mathbb{R}^d$. Each function f is defined by some $w \in \mathbb{R}^d$ as follows:

$$f(h) = \frac{w\cdot h}{||w||||h||}$$

OK, so the function isn't quite linear; we've normalized it.

In order to apply the PAC-Bayes methodology, we'll need to construct distributions on $\mathbb{R}^d$, which are a bit unfamiliar to us. We'll start with Π which is just $N(0,1)$ independently in each of the d coordinates. So far so good – but how could Λ be similarly simple? Thankfully, Π is rotationally symmetric, which will ultimately allow us to think in one dimension. The nice property about Π is that the (orthogonal) components of a sampled vector are independent. Since the normal distribution is rotationally symmetric in every direction, that nice property is preserved after a change of basis, as you saw in HW7. For any $\eta \in \mathbb{R}^d$, we can reorient the coordinate system to have one dimension (say, the first) parallel to η. To sample from Π, we still independently sample $N(0,1)$'s, but they correspond to different basis vectors. If we play our cards right, we can ignore all coordinates besides the first, since they are independent of the first coordinate, and are zero for η.

We design Λ to take advantage of spherical symmetry. It wil have parameters: a direction $\eta \in \mathbb{R}^d$ and a softener (to be explained later) $\mu > 0$. Thus, $\Lambda_{\mu,\eta}$ is $N(\mu, 1)$ in the direction of η and $N(0,1)$ in the perpendicular directions. By the previous reasoning, we can always 'align' Π with $\Lambda_{\mu,\eta}$. See how this simplifies the following calculation:

$$\begin{aligned} KL(\Lambda_{\mu,\eta} \| \Pi) =& \mathrm{KL}(N(\mu,1), N(0,1)) + \mathrm{KL}(\Lambda^{\perp}_{\mu,\eta}, \Pi^{\perp}) \\ =& \mu^2/2 + 0 \end{aligned}$$

The first equality follows from independence of the coordinates. The second equality follows from the aforementioned KL divergence for normals, and the congruence of $\Lambda^{\perp}_{\mu,\eta}$ and $\Pi^{\perp}$: both just sample $N(0,1)$'s for each coordinate.

Task

Prove that, for all distributions D, $\delta > 0$, with probability at least $1 - \delta$ over $S \sim D^m$: for all $\eta \in \mathbb{R}^d,\ \mu > 0$,

$$\mathrm{KL}(G_m(\Lambda_{\mu,\eta}), F(\Lambda_{\mu,\eta})) \leq \frac{\frac{\mu^2}{2} + \ln \frac{m+1}{\delta}}{m}$$

where the easy-to-compute empirical objective is

$$G_m(\Lambda_{\mu,\eta}) = \frac{1}{m} \sum_{i=1}^{m} (1 - \Phi(\mu f_i(\eta)))$$

where Φ is the normal CDF. This is easier to compute that $F_m(\Lambda_{\mu,\eta})$ since an invocation of the normal CDF replaces a multivariate integral. You should use the previous results, the parallel-perpendicular decomposition trick, and the normal distribution's symmetry and closure under linear combinations.

Oh, and congratulations – you're about to give a good bound for one of the most intensely studied and popularly used learning tasks. To explain this remark, I will (finally!) explain what μ is. Think about what happens as $\mu \to \infty$. If $f_i(\eta)$ is negative then the inner quantity $1 - \Phi(\mu f_i(\eta)$ approaches 1. If $f_i(\eta)$ is positive then the inner quantity approaches 0. In other words, this is some kind of approximation to classification. This approach to classification is very popular.

Problem 9: Stable stochastic optimization (extra credit) (10 pts.)

Let's consider stochastic optimization algorithms which aren't sensitive to small perturbations in the data S. Let:

- $S^{\backslash i} = S \backslash \{f_i\}$, i.e. the set of $m - 1$ samples where the ith sample is removed.
- $S^i = S^{\backslash i} \cup \{f'\}$ for some f' which we will specify later.

We will consider algorithms A which:

- take S and return a 'point' distribution which places probability 1 on a single $h \in H$.

- are β-**stable**: for all S and $i \in [m]$,

$$\max_f |f(A(S)) - f(A(S^{\setminus i}))| \leq \beta$$

By the triangle inequality, β-stability implies that for all S, f', and $i \in [m]$

$$\max_f |f(A(S)) - f(A(S^i))| \leq 2\beta$$

(Note that f' is hidden inside $S^{\setminus i}$.) Let $Z = F(A(S)) - F_m(A(S))$. It can be showed that $E(Z) \leq 2\beta$.

Task

Suppose $f(h) \in [0,\ C]$.

A. Let Z^i be Z when S is replaced by S^i. Prove $|Z - Z^i| \leq 4\beta + C/m$. Hint: use the triangle inequality.

B. Prove that with probability at least $1 - \delta$,

$$Z \leq 2\beta + m(4\beta + C/m)\sqrt{\frac{\log(1/\delta)}{2m}}.$$