Stat 991: Multivariate Analysis, Dimensionality Reduction, and Spectral Methods **Lecture: 7**

Dimensionality Reduction and Learning: Margins and Classification

Instructor: Sham Kakade

1 Preserving Inner Products

As a simple corollary, we see that inner products are preserved under random projection.

Corollary 1.1. *Let $u, v \in \mathbb{R}^d$ and that $\|u\| \leq 1$ and $\|v\| \leq 1$. Let $f = \frac{1}{\sqrt{k}} Ax$ where A is a $k \times d$ matrix, where each entry is sampled i.i.d from a Gaussian $N(0,1)$ (or from $U(-1,1)$). Then,*

$$\Pr(|u \cdot v - f(u) \cdot f(v)| \geq \epsilon) \leq 4e^{-(\epsilon^2 - \epsilon^3)k/4}$$

Proof. Applying Theorem **??** to the vectors $u+v$ and $u-v$, we have that with probability at least $1 - 4e^{-(\epsilon^2-\epsilon^3)k/4}$:

$$\begin{aligned} (1-\epsilon)\|u+v\|^2 \leq \quad \|f(u+v)\|^2 \quad \leq (1+\epsilon)\|u+v\|^2 \\ (1-\epsilon)\|u-v\|^2 \leq \quad \|f(u-v)\|^2 \quad \leq (1+\epsilon)\|u-v\|^2 \end{aligned}$$

Now we have:

$$\begin{aligned} 4f(u) \cdot f(v) &= \|f(u+v)\|^2 - \|f(u-v)\|^2 \\ &\geq (1-\epsilon)\|u+v\|^2 - (1+\epsilon)\|u-v\|^2 \\ &= 4u \cdot v - 2\epsilon(\|u\|^2 + \|v\|^2) \\ &\geq 4u \cdot v - 4\epsilon \end{aligned}$$

The proof of the other direction is analogous. □

2 Margin Based Classification

For now, assume we have a distribution over $X \in \mathcal{X} \subset \mathbb{R}^d$ and $Y \in \{-1, 1\}$. Assume that there exists a weight vector β such that $sign(\beta \cdot X) = Y$, with probability one. Hence, the distribution is separable. Furthermore, let us scale the distribution so that it is separable at margin 1, i.e.:

$$Y(\beta \cdot X) \geq 1$$

What learning algorithm should we use? The The VC dimension of halfspaces is $\Omega(D)$, so naively minimizing the $0/1$ loss in D dimensions may not be lead to good generalization properties (and it's not clear how to do this anyways). Instead, maximizing the margin can be shown to provide good generalization properties — however computationally, this may be a little cumbersome (even though it is polytime).

Let us say say we have a training set $T = \{(X_i, Y_i)\}_i$.

Often, what is done, is that the perceptron algorithm is run on the training set. The perceptron algorithm run on any sequence of points $\{(X_i, Y_i)\} \subset T$ sampled from this distribution makes at most:

$$M \leq \|\mathcal{X}\|^2 \|\beta\|^2$$

mistakes (regardless of the length of the sequence) where $\|\mathcal{X}\| = \max_{X \in \mathcal{X}} \|X\|$. Hence, if we repeatedly cycle through the dataset, then eventually we will no longer make mistakes.

But what about generalization? Naively using this perceptron predictor does not necessarily lead to good generalization behavior since the VC dimension of halfspaces is $\Omega(D)$ (and no bound is known for this convergent point of the perceptron).

2.1 Random Projections and Margin Preservation

Now let us project β and X by $P = \frac{1}{k}A$, where $A \in \mathbb{R}^{k \times d}$ and each entry in A is sample independently from $N(0,1)$. Is separability preserved under our training set?

Lemma 2.1. *Assume $\|\mathcal{X}\| \leq 1$. If $k = O(\|\beta\|^2 \|\mathcal{X}\|^2 \log \frac{n}{\delta})$, then with probability greater than $1 - \delta$ for all i*

$$P\beta \cdot PX_i \geq \frac{1}{2}$$

and

$$\frac{1}{2}\|X_i\|^2 \leq \|PX_i\|^2 \leq 2\|X_i\|^2, \quad \frac{1}{2}\|\beta\|^2 \leq \|P\beta\|^2 \leq 2\|\beta\|^2$$

Proof. Choose $\epsilon = \frac{1}{2\|\beta\|\|\mathcal{X}\|}$ and apply the inner product preserving lemma, which implies that for any particular X_i and β, that $|P\beta \cdot PX_i - \beta \cdot X_i| \geq \frac{1}{2}$, so that:

$$|Y_i P\beta \cdot PX_i - Y_i \beta \cdot X_i| \leq \frac{1}{2}$$

For $O(n^2)$ events, we use $O(\delta/n^2)$ so the total error probability is δ. The final claim follows from the norm preserving lemma. □

2.2 Generalization

If we run the perceptron algorithm, on the training set, then the total number of mistakes made is:

$$M_t \leq O(\|\beta\|^2 \|\mathcal{X}\|^2)$$

Note that this implies that after $O(\|\beta\|^2 \|\mathcal{X}\|^2)$ iteration the perceptron will stabilize to a constant solution, which has zero error.

For generalization, we are now working with a space of dimension $O(\|\beta\|^2 \log \frac{n}{\delta})$.

There are other methods to obtain generalization but the important point here is that under the margin assumption, we are essentially working in a finite dimensional space (and this subspace can be determine non-adaptively from the labels $\{Y_i\}$).

2.3 Random Projections and Maximum Likelihood Estimation

First note that if we project to $k = O(\frac{\log n}{\epsilon^2})$ dimensions then (using $P = \frac{1}{\sqrt{k}}A$), we have that for all i:

$$|P\beta \cdot PX_i - \beta \cdot X_i| \leq \|\beta\|\|X_i\|\epsilon$$

Let us define the loss using only $XP^\top$ as:

$$L_P(w) = \frac{1}{n}\mathbb{E}_Y\|Y - XP^\top w\|^2$$

Let β_P be the best fit of Y with $XP^\top$, i.e.

$$\beta_P = \arg\min_w L_P(w)$$

and let $\hat{\beta}_P$ be the MLE fit of Y with $XP^\top$ (so $\lambda = 0$). Now by the previous corollary, then:

$$\mathbb{E}_Y[L_P(\hat{\beta}_P)] - L_P(\beta_P) = \mathbb{E}_Y[\|XP^\top \hat{\beta}_P - XP^\top \beta_P\|^2] \leq \frac{k}{n}$$

Also note that:

$$\begin{aligned}
L_P(\beta_P) &\leq L_P(P\beta) \\
&= \frac{1}{n}\mathbb{E}[\|Y - XP^\top P\beta\|^2] \\
&= \frac{1}{n}\mathbb{E}[\|Y - X\beta\|^2] + \frac{1}{n}\|X\beta - XP^\top P\beta\|^2 \\
&= L(\beta) + \frac{1}{n}\sum_i (P\beta \cdot PX_i - \beta \cdot X_i)^2 \\
&\leq L(\beta) - \|\beta\|^2(\frac{1}{n}\sum_i \|X_i\|^2)\epsilon^2 \\
&= L(\beta) - \|\beta\|^2\|\Sigma\|_{\text{trace}}\epsilon^2
\end{aligned}$$

Theorem 2.2. *(Risk Bound after Random Projection) Assuming* $\mathrm{Var}(Y_i) \leq 1$*, and that* P *is* ϵ *inner product preserving for* $k = O(\frac{\log n}{\epsilon^2})$*, then:*

$$\mathbb{E}_Y\|XP^\top\hat{\beta}_P - X\beta\|^2 = \mathbb{E}_Y[L_P(\hat{\beta}_P)] - L(\beta) \leq \frac{k}{n} + \|\beta\|^2\|\Sigma\|_{trace}^2\epsilon^2 = \frac{20\log n}{n\epsilon^2} + \|\beta\|^2\|\Sigma\|_{trace}\epsilon^2$$

Hence, choosing $\epsilon^2 = O(\sqrt{\frac{\log n}{n\|\beta\|^2\|\Sigma\|_{trace}}})$*, implies that* $k = O(\|\beta\|\sqrt{\|\Sigma\|_{trace} n\log n})$ *and:*

$$\mathbb{E}_Y\|XP^\top\hat{\beta}_P - X\beta\|^2 \leq O\left(\frac{\sqrt{\log n\|\beta\|^2\|\Sigma\|_{trace}}}{\sqrt{n}}\right)$$

Proof. From above we have that:

$$L(\beta) \geq L_P(\beta_P) - \|\beta\|^2\|\Sigma\|_{\text{trace}}\epsilon^2$$

so that:

$$\mathbb{E}_Y[L_P(\hat{\beta}_P)] - L(\beta) \leq \mathbb{E}_Y[L_P(\hat{\beta}_P)] - L_P(\beta_P) + \|\beta\|^2\|\Sigma\|_{\text{trace}}\epsilon^2 = \mathbb{E}_Y[\|XP^\top\hat{\beta}_P - XP^\top\beta_P\|^2] + \|\beta\|^2\|\Sigma\|_{\text{trace}}\epsilon^2$$

and we have bounded the risk in the last terms as $\frac{k}{n}$. □

This matches the risk bound up to log factors. Also, our algorithm is simply an MLE estimate in $k = O(\|\beta\|\sqrt{\|\Sigma\|_{\text{trace}} n\log n})$ dimensions. Note that the number of dimensions we choose is growing as $O(\sqrt{n})$.