SML Recitation Notes Week 9: Low-Noise Condition and Random Projection

Min Xu

April 3, 2011

1 Low-Noise Condition

1.1 ERM vs. Plug-in

Empirical Risk Minimization (ERM) approach to learning classifier is to

1. Describe a class of classification functions $\mathcal{H}$ (also called hypothesis class)
2. Define a notion of loss, either 0-1 loss or a surrogate loss
3. Given the training data, find classifier $h \in \mathcal{H}$ by minimizing average loss among training data (Empirical Risk)

We evaluate ERM by comparing

$$\mathbb{E}_{\text{samples}}[R(\hat{h}) - R(h^*)]$$

where $h^* = \min_{h \in \mathcal{H}} R(h)$ is the optimal classifier in the class, and $R(h) = P(Y \neq h(X))$. This evaluation is unfair in that we are only comparing functions within the same class.

Examples include Logistic regression, SVM, decision tree, etc. Note that the learning algorithm itself might not explicitly minimize empirical risk; for example, SVM appears to be maximizing the margin although we can show that it is equivalent to minimizing the average hinge loss.

Plug-in approach to learning classifier is to

1. Estimate conditional densities $p(X|Y = 1)$ and $p(X|Y = 0)$ with kernels, produce estimate $\hat{\eta}(X) = \hat{p}(Y = 1|X)$
2. Output classifier $\hat{f}(X) = \mathbb{I}_{\hat{\eta}(X) \geq 1/2}(X)$

We evaluate Plug-in classifier by comparing

$$\mathbb{E}_{\text{samples}}[R(\hat{f}) - R(f^*)]$$

where f^* is the Bayes optimal estimator: $f^*(X) = 1$ iff $\eta(X) = P(Y = 1|X) \geq 1/2$. Here we compare plug-in $\hat{f}$ against the optimal among all classifiers, but we need assumptions on the true distribution.

Examples include nearest-neighbors, local-polynomial regression. Note that when we learning plug-in classifier, we do not explicitly compute density estimates either, e.g. nearest-neighbor.

Remark:

- ERM need not suffer in high-dimensions. For example, if the high-dimensional data is linearly separable, then SVM or logistic regression will converge to a correct linear classifier at the same rate as for low-dimensional data (at rate $O(\sqrt{\frac{1}{n}})$ by Vapnik-Chernonenkis)

- Plug-in classifier could suffer in high-dimensions. Suppose we have features $X_1, ...X_d$ but a lot of the features are irrelevant, then intuitively, that could throw off nearest-neighbor classification.
  More precisely, analysis on nonparametric regression show that the convergence rate of $\hat{\eta}$ to true η is $O(n^{-\frac{\beta}{2\beta+d}})$ (where β is smoothness of the true η) The dependency on d comes from the variance of KDE: larger dimensions exponentially enlarges the space and hence exponentially drives up the variance.

1.2 Low-Noise Condition

Yet plug-in classifiers often do very well in high-dimensions, for example, nearest-neighbor classification is very effective in computer vision. Intuitively, we expect plug-in classifier to do well when there aren't very irrelevant features; the Low-Noise condition formalizes this intuition:

Definition 1. (Low-Noise Condition)
Let P_{XY} be a distribution on $\mathcal{X} \times \mathcal{Y}$. Then the distribution satisfies low noise condition if for constant $C_0 > 0$ and $\alpha \geq 0$, and for all t, we have

$$P_X(0 < |\eta(X) - 1/2| \leq t) \leq C_0 t^\alpha$$

Remark:

- The larger the α, the lower the noise. As we will see later, the larger the dimension, the larger α must be in order for Plug-in classifier to maintain good performance.

- This is a very general condition that covers more cases than just "irrelevant features". But let us see how our intuition about irrelevant features fit in to the Low-noise Condition: if features $X_1...X_l$ are irrelevant, $X_1, ..., X_l$ do not predict Y at all, so $P(Y = 1|X_1, ..., X_l) \approx \frac{1}{2}$.

  Hence, by including features $X_1, ..., X_l$, we increased dimension d but did not increase α, leading to overall decrease in classifier performance.

1.3 Audibert and Tsybakov's Analysis

Audibert and Tsybakov analyzed (almost) local polynomial regression.

Remark: Recall that in local polynomial regression, we evaluate regression function $\hat{m}$ at point x by fitting a polynomial centered at x:

$$\min_{a_0,...,a_p} \sum_{i=1}^{n} K(x, X_i)(Y_i - (a_0 + a_1(X_i - x) + ... + a_p(X_i - x)^p))^2$$

The above minimization has a closed form. We then discard $a_1, ..., a_p$ and take $\hat{m}(x) = a_0$

Theorem 2. *(Audibert and Tsybakov) Assuming*

- *For all x, with probability around $\exp(-a_n\delta^2)$, $|\hat{\eta}(x) - \eta(x)| \leq \delta$ (a_n is some increasing function of n, depend on β)*

- *Low-noise condition on true distribution*
- β*-smoothness on true regression function* $\eta(x)$

Then with bandwidth $h = n^{-\frac{1}{2\beta+d}}$*, we get*

$$\sup_{p \in \mathcal{P}} \mathbb{E}[R(\hat{f}) - R(f^*)] \leq Cn^{-\frac{\beta(1+\alpha)}{2\beta+d}}$$

Here we see that if $\beta\alpha > d$, then the rate will not decrease with d. If β-smoothness is fixed, then as dimension d increases, we will need lower and lower noise (higher and higher α)

We will sketch Audibert and Tsybakov's argument, highlighting key steps and omitting technical derivation.

Proof. (Sketch)
The key steps:

1. Apply two important identities:

   $$\mathbb{E}[R(\hat{f}) - R(f^*)] = \mathbb{E}[|2\eta(X) - 1|\mathbb{I}_{\hat{f}(X) \neq f^*(X)}]$$
   $$\mathbb{I}_{\hat{f}(X) \neq f^*(X)} \leq \mathbb{I}_{|\eta(X) - \frac{1}{2}| \leq |\hat{\eta}(X) - \eta(X)|}$$

   The second identity follows because the event $\hat{f}(X) \neq f^*(X)$ implies that $|\eta(X) - \frac{1}{2}| \leq |\hat{\eta}(X) - \eta(X)|$

   Combining these gives us:

   $$\mathbb{E}[R(\hat{f}) - R(f^*)] \leq \mathbb{E}[|2\eta(X) - 1|\mathbb{I}_{|\eta(X) - \frac{1}{2}| \leq |\hat{\eta}(X) - \eta(X)|}]$$

   Which relates the excess risk to two quantities: $|\eta(X) - \frac{1}{2}|$ and $|\hat{\eta}(X) - \eta(X)|$

2. Divide feature space $\mathcal{X} = \mathbb{R}^d$ into regions:

   $A_0 = \{x \in \mathbb{R}^d : 0 < |\eta(x) - \frac{1}{2}| \leq \delta\}$

   ...

   $A_j = \{x \in \mathbb{R}^d : 2^{j-1}\delta < |\eta(x) - \frac{1}{2}| \leq 2^j\delta\}$ Intuitively, A_0 is hard and for large j, A_j is easy

   Note that the risk decomposes into risks for each region:

   $$\mathbb{E}[|2\eta(X) - 1|\mathbb{I}_{|\eta(X) - \frac{1}{2}| \leq |\hat{\eta}(X) - \eta(X)|}] = \sum_{j=0}^{\infty} P(X \in A_j)\mathbb{E}[|2\eta(X) - 1|\mathbb{I}_{|\eta(X) - \frac{1}{2}| \leq |\hat{\eta}(X) - \eta(X)|}|X \in A_j]$$

3. We analyze excess risk for each region separately.

   On hard regions (A_j for small j), cost $|\eta(X) - \frac{1}{2}|$ is low but we have to pay that cost very often because $\hat{\eta}(X) - \eta(X) \geq |\eta(X) - \frac{1}{2}|$ happens very often.

   On easy regions (A_j for large j), cost $|\eta(X) - \frac{1}{2}|$ is high but we almost never have to pay that cost because $\hat{\eta}(X) - \eta(X) \geq |\eta(X) - \frac{1}{2}|$ will never happen. Hence, the hard regions contribute a lot more to the overall excess risk.

4. However, by the low noise condition, the hard regions occur with low probability, and hence, the overall excess risk averaged across all regions is still low.

□

2 Random Projection

2.1 All about Projections

Definition 3. Let $S \subset \mathbb{R}^p$ be a set, let $v \in \mathbb{R}^p$ be a vector, then the orthogonal projection of v onto s is a vector $w^* \in \S$ such that:

$$Proj_S(v) = w^* = \arg\min_{w \in S} ||w - v||_2$$

Suppose now that S is a hyperplane, then we can compute the projection of v onto S in closed form. The following two procedures both compute the projection onto a hyperplane and are equivalent:

Let S have dimension k

1. (Orthonormal Basis) Let $w'_1, ..., w'_k$ be an **orthonormal** basis for S. Let $W \in \mathbb{R}^{p \times k}$ be a basis matrix such that the i-th column of W is w'_i.
   Then $Proj_S(v) = WW^\mathsf{T} v$
2. (Basis) Let $w_1, ..., w_k$ be a basis for S, let $W \in \mathbb{R}^{p \times k}$ be a basis matrix such that the i-th column of W is w_i.
   Then $Proj_S(v) = W(W^\mathsf{T} W)^{-1} W^\mathsf{T} v$

In orthonormal basis case, $W^\mathsf{T} v \in \mathbb{R}^k$ is the vector of coefficients for the projection; it is in a lower dimensional space and often people will abuse vocabulary and say that $W^\mathsf{T} v$ is the projection.

2.2 Random Projection

Lemma 4. *(Sampling Lemma)*
Let $x \in \mathbb{R}^p$ be a random unit vector, then x is sampled uniformly from the unit sphere S^p if and only if x is distributed identically to some random vector $\frac{X}{||X||_2}$ where $X = (X_1, ..., X_p) \sim N(0, Id_p)$

Proof. The density of $N(0, \mathrm{Id}_p)$ is radially symmetric. □

Definition 5. (Real Random Projection) Let $v \in \mathbb{R}^p$. Draw unit vectors $x_1, ..., x_k$ uniformly at random. Form $p \times k$ matrix $X = (x_1, ..., x_k)$ and the random projection of v is $v' = X(X^\mathsf{T} X)^{-1} X^\mathsf{T} v$ where $(X^\mathsf{T} X)^{-1} X^\mathsf{T} v$ is the lower dimensional coefficient vector.

We say that the subspace spanned by $(x_1, ..., x_k)$ is a *random subspace*.

Definition 6. (Fake Random Projection) Let $v \in \mathbb{R}^p$. Draw unit vectors $x_1, ..., x_k$ uniformly at random. Form $p \in \times k$ matrix $X = (x_1, ..., x_k)$ and the random projection of v is $v' = XX^\mathsf{T} v$ where $X^\mathsf{T} v$ is the lower dimensional coefficient vector.

Theorem 7. *(Johnson-Lindenstrauss) Let $\{v_i\}_{i=1}^n$ be a set of n points in $\mathbb{R}^p$. Let $0 < \epsilon \leq 1$, let $k \geq \frac{\log n}{\epsilon^2}$, then with probability $\frac{1}{n}$, for all pairs i, j,*

$$(1 - \epsilon)||v_i - v_j||_2 \frac{k}{d} \leq ||\Pi_k(v_i) - \Pi_k(v_j)|| \leq (1 + \epsilon||v_i - v_j||_2 \frac{k}{d}$$

Where $\Pi_k(v_i)$ is the random projection (can be real or fake) of v_i onto k-dimensional space.

Preserving pair-wise distance is useful property; for instance, it allows us to do nearest neighbor search on a much lower dimensional space.

We can achieve the same distance-preserving dimensionality reduction through PCA, but that is much more computationally expensive.

To get some more intuition about the random projection, we state a few more facts:

Proposition 8. *The following two procedures are equivalent and both produce a random subspace:*

- *Draw unit vectors* $(x_1, ..., x_k)$ *uniformly at random. Take subspace spanned by* $(x_1, ..., x_p)$.
- *Draw random unit vector* x_1' *from entire sphere* S^p. *Let* S^{p-1} *be the lower dimensional sphere orthogonal to* x_1, *draw unit vector* x_2' *randomly from* S^{p-1}. *Repeat. We get orthonormal vectors* $x_1', ..., x_k'$ *as basis for a subspace.*