Sobolev Models Homework

Machine Learning Theory

1 Introduction

This week, we'll bound the gaussian width of a neighborhood in a Sobolev model. This one.

$$\mathcal{M}^p = \left\{ m(x) = \sum_{j=0}^{\infty} m_j \phi_j(x) \ : \ \rho_{\text{-}\Delta^p}(m) \le B \right\}$$
$$\text{where } \rho_{\text{-}\Delta^p}\left(\sum_{j=0}^{\infty} m_j \phi_j \right) = \sqrt{\sum_{j=0}^{\infty} \lambda_j^p m_j^2 \le B^2} \tag{1}$$

$$\text{for } \phi_j(x) = \sqrt{2}\cos(\pi j x) \quad \text{and} \quad \lambda_j = \pi^2 j^2.$$

We'll use this to bound the error $\|\hat{\mu} - \mu\|_{L_2(P)}$ of this least squares estimator.

$$\hat{\mu} = \operatorname*{argmin}_{m \in \mathcal{M}^p} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - m(X_i)\}^2 \tag{2}$$

2 The Simplified Case

2.1 The whole thing.

We'll start with a simple version of the problem. We'll bound the gaussian width of this Sobolev-like model. The whole thing—not a neighborhood in it.

$$\mathcal{M} = \left\{ m(x) = \sum_{j=0}^{\infty} m_j \phi_j(x) : \sum_{j=0}^{\infty} \lambda_j m_j^2 \le B^2 \right\}$$
$$\text{where } \langle \phi_j, \phi_j \rangle_{L_2(P)} = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The Sobolev model described in (1) is a model like this when our covariate X_i has a uniform distribution on $[0, 1]$, as in that case the inner product $\langle u, v \rangle_{L_2(P)}$ agrees with the inner product $\langle u, v \rangle_{L_2}$ for which our cosine basis is orthonormal.

Rather than bounding gaussian width itself, we'll bound a related quantity.

$$\mathrm{w}_2(\mathcal{M}) = \sqrt{\mathrm{E}\max_{m\in\mathcal{M}}\langle g, m\rangle_{L_2(\mathrm{P_n})}^2} \quad \text{where} \quad g_i \overset{iid}{\sim} N(0,1).$$

We know this is *larger* than gaussian width. In terms of $Z = \max_{m\in\mathcal{M}}\langle g, m\rangle_{L_2(\mathrm{P_n})}$,

$$\mathrm{w}(\mathcal{M}) = \mathrm{E}\, Z \quad \text{and} \quad \mathrm{w}_2(\mathcal{M}) = \sqrt{\mathrm{E}\, Z^2} = \sqrt{(\mathrm{E}\, Z)^2 + \mathrm{Var}(Z)}.$$

Exercise 1 *Bound the gaussian width of the model* (3)*. To do this, bound*

$$\mathrm{w}_2(\mathcal{M}) = \sqrt{\mathrm{E}\left(\max_{m\in\mathcal{M}}\frac{1}{n}\sum_{i=1}^n g_i\left\{\sum_{j=0}^\infty m_j\phi_j(X_i)\right\}\right)^2}.$$

Your bound should depend on the sequence λ_j.

Hint. Reordering the sums, what we get is a dot product between the sequence of coefficients m_j and a gaussian-weighted average of the evalauted basis functions $\phi_j(X_i)$. That is,

$$\sum_{i=1}^n g_i\left\{\sum_{j=0}^\infty m_j\phi_j(X_i)\right\} = \sum_{j=0}^\infty\left\{\sum_{i=1}^n g_i\phi_j(X_i)\right\}m_j$$

What does the Cauchy-Schwarz inequality for the inner product $\langle u, v\rangle = \sum_{j=0}^\infty u_j v_j$ and the constraint $\sum_j \lambda_j m_j^2 \le B^2$ tell us about this? It helps to multiply each term by a fancy version of 1: $\sqrt{\lambda_j}/\sqrt{\lambda_j} = 1$.

2.2 A Neighborhood

Now let's think about a neighborhood of zero in this model. Because our basis is orthonormal, it's very easy to express our neighborhood constraint in terms of the coefficients m_j.

$$\begin{aligned}\mathcal{M}_s &= \{m \in \mathcal{M} \ : \ \|m\|_{L_2(P)} \le s\} \\ &= \left\{m(x) = \sum_{j=0}^\infty m_j\phi_j(x) : \ \sum_{j=0}^\infty \lambda_j m_j^2/B^2 \le 1 \quad \text{and} \quad \sum_{j=0}^\infty m_j^2/s^2 \le 1\right\}.\end{aligned} \tag{4}$$

Exercise 2 *Explain why the two lines of* (4) *describe the same set.*

Now we're ready to bound the width of this neighborhood.

Exercise 3 *Bound the gaussian width of this neighborhood $\mathcal{M}_s$. Your bound should depend on the radius s and the sequence λ_j.*

Tip. You should be able to lean on your argument from Exercise 1. Note that if a sequence of coefficients m_j satisfies the constraints $\sum_j a_j m_j^2 \le 1$ and $\sum_j b_j m_j^2 \le 1$, it satisfies the sum of those constraints, $\sum_j (a_j + b_j) m_j^2 \le 2$.

2.3 Error Bounds for Least Squares

Now let's use this to calculate a high probability bound on the error of the least squares estimator.

Exercise 4 *Calculate a bound of the form $\|\hat{\mu} - \mu^\star\|_{L_2(P)} \le s$ for the least squares estimator $\hat{\mu}$ that holds with high probability when $\mu = 0$. First, express this in terms of the sequence λ_j. Then, assuming that X_i is uniformly distributed on $[0,1]$, do it specifically for the model (1) and simplify as much as possible. Try to get a bound proportional to $n^{-\beta}$ for β depending on p.*

Tip. When bounding sums like $\sum_{j=0}^{\infty} f(j)$, it's often helpful to use the integral approximation $\int_0^\infty f(x)dx$.

3 Getting Practical Conclusions

In Exercise 4, we established a bound on the error of the least squares estimator in (1) when some simplifying assumptions are satisfied. It's almost useful, but our simplifying assumptions really limit its applicability.

1. It's valid only when $\mu = 0$. That's because our approach to bounding $\|\hat{\mu} - \mu^\star\|_{L_2(P)}$ depends on the gaussian width of a neighborhood of $\mu^\star$. We've only bounded the width of a neighborhood of zero, so our bound works only when $\mu^\star = 0$.

2. It's valid only when X_i is uniformly distributed on $[0,1]$. That's because we've assumed that our basis functions $\phi_0, \phi_1, \ldots$ are orthonormal for the inner product $\langle u, v\rangle_{L_2(P)}$. The cosine basis functions from (1) are orthonormal for the inner product $\langle u, v\rangle_{L_2}$, which is the same as $\langle u, v\rangle_{L_2(P)}$ if and only if X_i has that uniform distribution.

These are both easy to fix. What we need to do is bound the gaussian width of the neighborhood $\mathcal{M}_s = \{m \ : \ \rho_{\text{-}\Delta^p}(m) \le B \ \text{ and } \ \|m - \mu^\star\|_{L_2(P)} \le s\}$. Or, because gaussian width is translation invariant, the width of the centered neighborhood $\mathcal{M}_s - \mu^\star$. And because $\mathrm{w}(\mathcal{V}) \le \mathrm{w}(\mathcal{V}^+)$ if $\mathcal{V} \subseteq \mathcal{V}^+$ for any curve m, we can use our result from Exercise 4 to do this if we can show that $\mathcal{M}_s - \mu^\star$ is contained in a set $\mathcal{M}_s^+$ like the neighborhoods we'd consider under our simplifying assumptions. That is, we can use that result if we can show that for some budget B_+ and radius s_+,

$$\mathcal{M}_s - \mu^\star \subseteq \mathcal{M}_s^+ \quad \text{where} \quad \mathcal{M}_s^+ = \{m_+ \in \mathcal{M}^p \ : \ \rho_{\text{-}\Delta^p}(m_+) \le B_+ \ \text{ and } \ \|m_+\|_{L_2} \le s_+\}.$$

Exercise 5
This one is optional.

1. *Show that if our covariates X_i are in $[0,1]$ with a probability density function f_X, then $\mathcal{M}_s - \mu^\star \subseteq \mathcal{M}_s^+$ for $B_+ = 2B$ and $s_+ = s\sqrt{M}$ for $M = \max_{x\in[0,1]} f_X(x)$.*

2. *Use this to bound* $\mathrm{w}(\mathcal{M}_s)$ *by* a multiple of *the width we calculated under our simplifying assumptions. That is, find α so that* $\mathrm{w}(\mathcal{M}_s) \le \alpha\, \mathrm{w}(\mathcal{M}_s^{simple})$ *where $\mathcal{M}_s^{simple} = \{m \ : \ \rho_{\text{-}\Delta^p}(m) \le B \ \text{ and } \ \|m\|_{L_2} \le s\}$.*

3. *Calculate a bound of the form $\|\hat{\mu} - \mu^\star\|_{L_2(P)} \le s$ for the least squares estimator $\hat{\mu}$ that holds with high probability. How much bigger is it than the bound you calculated in Exercise 4?*

Hint. $\mathrm{w}(\mathcal{M}_s) = \mathrm{w}(\mathcal{M}_s - \mu^\star)$ and $\rho(m - \mu^\star) \le \rho(m) + \rho(\mu^\star)$.

Hint. Review the stuff on generalization at the beginning of the population least squares lecture.

Hint. $\mathrm{w}(\alpha\mathcal{V}) = \alpha\, \mathrm{w}(\mathcal{V})$ for $\alpha\mathcal{V} = \{\alpha v \ : \ v \in \mathcal{V}\}$.