19. Linear Predictors (part 2: regression)

Monday, February 19, 2024 5:01 PM

§9.2 Linear Regression

$\mathcal{H} = L_d$, $X \subseteq \mathbb{R}^d$, $Y = \mathbb{R}$. Usually use squared loss $\ell(h,(x,y)) := (h(x)-y)^2$

using ℓ^1 loss $|h(x)-y|$ is a LP

so $\hat{L}_S(h) = \frac{1}{m}\sum_{i=1}^m (h(x_i)-y_i)^2$ is mean-squared-error, MSE. Very traditional.

ERM is least-squares problem

$$\underbrace{\operatorname*{argmin}_{w\in\mathbb{R}^d} \frac{1}{m}\sum_{i=1}^m (\langle w, x_i\rangle - y_i)^2}_{= \frac{1}{m}\|X\cdot w - y\|_2^2} \qquad \text{let } X = \begin{array}{c} \leftarrow d \rightarrow \\ \begin{bmatrix} x_1^T \\ \vdots \\ x_m^T \end{bmatrix} \end{array} \begin{array}{c} \uparrow \\ m \\ \downarrow \end{array}$$

(Assume $m \geq d$ for now)

prefer matrix/vector format numerically

solve in 1000's of ways.

Some major classes of sol'n method:

① dense methods.

Solution satisfies normal equations,

$$X^TX = \sum_{i=1}^m x_i x_i^T \qquad X^T(Xw - y) = 0, \text{ i.e. } w = (X^TX)^{-1}X^Ty$$

Do not call inv(X^TX)! Don't do cholesky on X^TX either! (slightly better) } these are inaccurate though fast

Instead: QR

$$m\,\overset{d}{\boxed{X}} = \boxed{Q}\boxed{R}$$

$Q^TQ = I$, R is upper triangular. eg. Gram-Schmidt, $O(md^2)$

*even better: QR w/ pivoting

(see code demo on github)

$$w = (X^TX)^{-1}X^Ty = (R^T\cancel{Q^TQ}R)^{-1}R^TQ^Ty = (R^TR)^{-1}R^TQ^Ty$$

cancel by hand

$$= R^{-1}\cancel{R^{-T}R^T}Q^Ty$$

cancel by hand

and solve $w = R^{-1}Q^Ty$ as $Rw = Q^Ty$

using backward substitution since R is triangular $O(d^2)$ time instead of $O(d^3)$

(Matlab's backslash $w = X\backslash y$ will do a good job)

② If large, use Krylov subspace method like conjugate gradient (CG) or better yet a variant optimized for least squares (LSQR...)

Can be combined w/ preconditioners

Always beats gradient descent

③ Stochastic Gradient Descent (SGD) if huge

④ New state-of-the art methods (randomized, hybrid, approximate, ...) if extremely large and/or very ill-conditioned.

19a. Linear Predictors (part 2: regression)

Monday, February 19, 2024 5:16 PM

What if I want to do quadratic regression or other polynomial regression?

① Sure... but it can be recast as linear regression by adding features!

e.g. $X = \mathbb{R}$, $p(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3$

embed $x \mapsto \varphi(x) = (1, x, x^2, x^3) \in \mathbb{R}^4$ and do (homogeneous) linear classification here.

e.g. $X = \mathbb{R}^2$, $p(x) = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_1 x_2 + a_4 x_1^2 + a_5 x_2^2$

$\varphi(x) = (1, x_1, x_2, x_1 \cdot x_2, x_1^2, x_2^2)$

② In high d, think twice... it's a lot of parameters

Sometimes we tame this by not considering cross-terms.

Learning theory for linear regression (ch. 11 in Mohri)

via Rademacher complexity

generic result: Thm 11.3 If the range of the loss ℓ is bounded in $[0, M]$ and $\hat{y} \mapsto \ell(\hat{y}, y)$ is μ-Lipschitz continuous, then

$$L_D(h) \le \hat{L}_S(h) + \begin{cases} 2\mu R_m(\mathcal{H}) + M \cdot \sqrt{\log(1/\delta)/2m} \\ 2\mu \hat{R}_S(\mathcal{H}) + 3M\sqrt{\log(2/\delta)/2m} \end{cases}$$

The problem?

If $\ell(\hat{y}, y) = (\hat{y} - y)^2$, this is neither bounded nor Lipschitz unless y and $\hat{y}$ $(= \langle w, x \rangle)$ are bounded, which is an unusual (though not unheard of) assumption. e.g., bound $\|w\|$ (as we saw in HW 3) and assume $\|x\|$ is bounded.

Alternatives: pseudo-dimension (§11.2.3 Mohri)

"Shattering" (and hence VC dim.) made sense when Y was finite, but if $Y = \mathbb{R}$, what to do?

19b. Linear Predictors (part 2: regression)

Wednesday, February 21, 2024 2:38 PM

a new notion of "shattering"

Def 11.4 Shattering of real-valued function families

Let $\mathcal{F}$ be a family of functions $f: Z \to \mathbb{R}$ (e.g. $Z = X \times Y$, $f = \ell \circ h$, $h \in \mathcal{H}$)

A set $C = (z_1, \ldots, z_m) \subseteq Z$ is shattered by $\mathcal{F}$ if $\exists\ (t_1, \ldots, t_m) \subseteq \mathbb{R}^m$ ← "witnesses" to the shattering

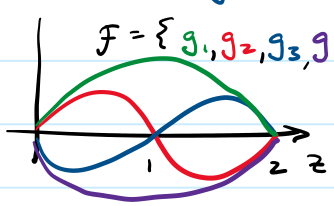
s.t.

$$\left| \left\{ \begin{bmatrix} \text{sign}(g(z_1) - t_1) \\ \vdots \\ \text{sign}(g(z_m) - t_m) \end{bmatrix} : g \in \mathcal{F} \right\} \right| = 2^m$$

↖ cardinality, not abs. value

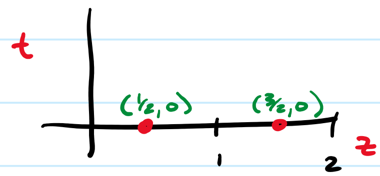
You pick $z_1, \ldots, z_m$ and $t_1, \ldots, t_m$ once, then look at all $g \in \mathcal{F}$

Ex $\mathcal{F} = \{g_1, g_2, g_3, g_4\}$



$z_1 = 1/2$, $t_1 = 0$

then $z_2 = 3/2$, $t_2 = 0$ are shattered



Def 11.5 The pseudo-dimension of $\mathcal{F}$ is the size of the largest set that can be shattered by $\mathcal{F}$.

aka "pdim"

Facts (see Mohri)

- $\text{pdim}(L_d) = d+1$ $(= \text{VCdim}(\text{sign} \circ L_d))$ via reducing to VCdim w/ thresholds

- There are generalization error bounds involving pdim (... but we won't cover. Not as nice/natural as VCdim)