Using Unlabeled Data for Supervised Learning

10-701

15-781

Tom Mitchell

November 2002

Setting:

- X set of instances governed by unknown $P(X)$
- $f : X \to Y$ target function (or, $P(Y|X)$)
- H set of possible hypotheses (models of f)

Given:

- iid labeled examples $L = \{\langle x_1, y_1\rangle, \ldots \langle x_m, y_m\rangle\}$
- iid unlabeled examples $U = \{x_{m+1}, \ldots x_{m+n}\}$

Determine:

$$\hat{f} \leftarrow \operatorname*{argmin}_{h \in H} \Pr_{x \in P(X)}[h(x) \neq f(x)]$$

Question:

- Under what conditions can U help?

When Can U Help?

Simple answer:

- If x_i in $L = \{\langle x_1, y_1 \rangle, \ldots \langle x_m, y_m \rangle\}$ are incompletely observed
- Then use $U \rightarrow \hat{P}(X)$ to complete x_i

What if x_i are fully observed?

What if x_i are fully observed?

Can use $U \rightarrow \hat{P}(X)$ to alter optimization problem

- Wish to find

$$\hat{f} \leftarrow \underset{h \in H}{\operatorname{argmin}} \sum_{x \in X} \delta(h(x) \neq f(x)) P(x)$$

- Often approximate as

$$\hat{f} \leftarrow \underset{h \in H}{\operatorname{argmin}} \frac{1}{|L|} \sum_{\langle x, y \rangle \in L} \delta(h(x) \neq y)$$

  and when $y = f(x)$, this is just

$$\hat{f} \leftarrow \underset{h \in H}{\operatorname{argmin}} \sum_{x \in X} \delta(h(x) \neq f(x)) \frac{n(x, L)}{|L|}$$

- Can use U for improved approximation:

$$\hat{f} \leftarrow \underset{h \in H}{\operatorname{argmin}} \sum_{x \in X} \delta(h(x) \neq f(x)) \frac{n(x, L) + n(x, U)}{|L| + |U|}$$

Idea: Use U to Avoid Overfitting

[Schuurmans 1997]

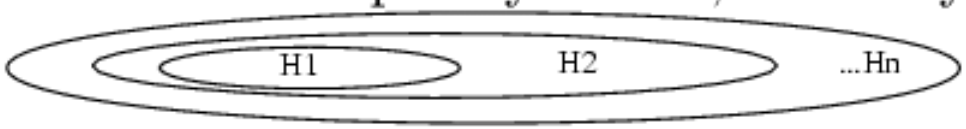
Define *metric* over $H \cup \{f\}$

$$d(h_1, h_2) \equiv \int \delta(h_1(x) \neq h_2(x)) p(x) dx$$

$$\hat{d}(h_1, f) = \frac{1}{|L|} \sum_{x_i \in L} \delta(h_1(x_i) \neq y_i)$$

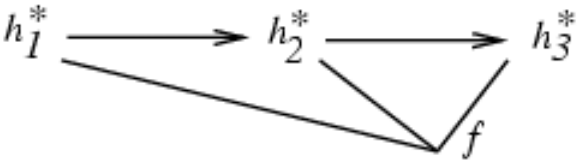
$$\hat{d}(h_1, h_2) = \frac{1}{|U|} \sum_{x \in U} \delta(h_1(x) \neq h_2(x))$$

Organize H into complexity classes, sorted by $P(h)$

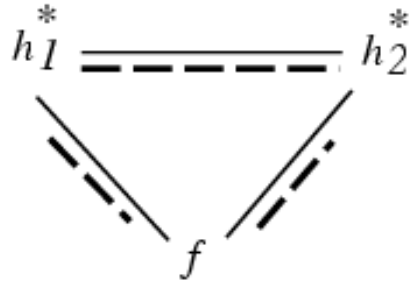


Let h_i^* be hypothesis with lowest $\hat{d}(h, f)$ in H_i

Prefer h_1^*, h_2^*, or h_3^*?



Idea: Use U to Avoid Overfitting



Note:

- $\hat{d}(h_i^*, f)$ optimistically biased (too short)
- $\hat{d}(h_i^*, h_j^*)$ unbiased
- Distances must obey triangle inequality!

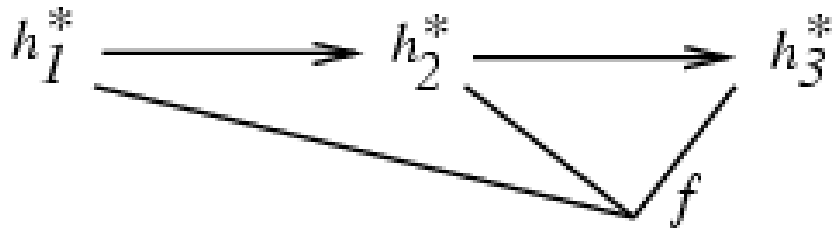
$$d(h_1, h_2) \leq d(h_1, f) + d(f, h_2)$$

$\rightarrow$ Heuristic:

- Continue training until $\hat{d}(h_i, h_{i+1})$ fails to satisfy triangle inequality

Procedure TRI

- Given hypothesis sequence $h_0, h_1, ...$
- Choose the last hypothesis h_ℓ in the sequence that satisfies the triangle inequality $d(h_k, h_\ell) \leq d(h_k, \widehat{\mathrm{P}_{Y|X}}) + d(h_\ell, \widehat{\mathrm{P}_{Y|X}})$ with every preceding hypothesis h_k, $0 \leq k < \ell$. (Note that the inter-hypothesis distances $d(h_k, h_\ell)$ are measured on the *un*labeled training data.)



Experimental Evaluation of TRI

[Schuurmans & Southey, MLJ 2002]

- Use it to select degree of polynomial for regression
- Compare to alternatives such as cross validation, structural risk minimization, …

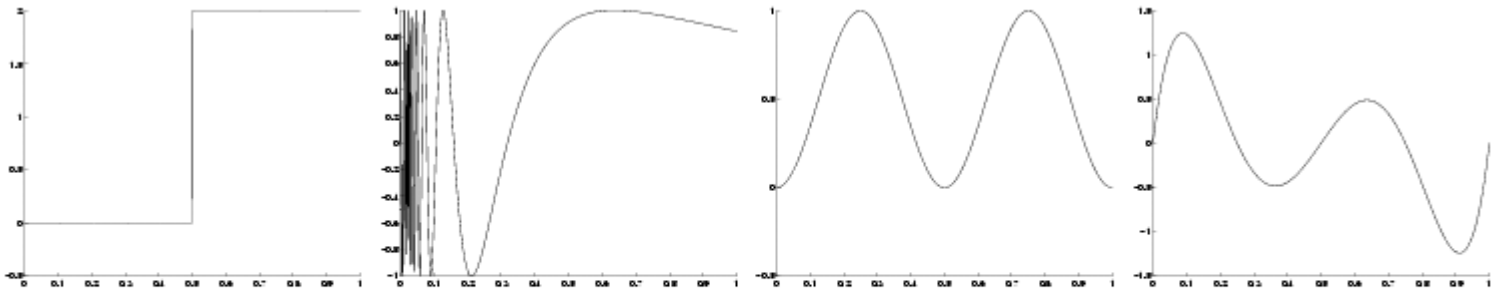


Figure 5: Target functions used in the polynomial curve fitting experiments (in order): $\text{step}(x \geq 0.5)$, $\sin(1/x)$, $\sin^2(2\pi x)$, and a fifth degree polynomial.

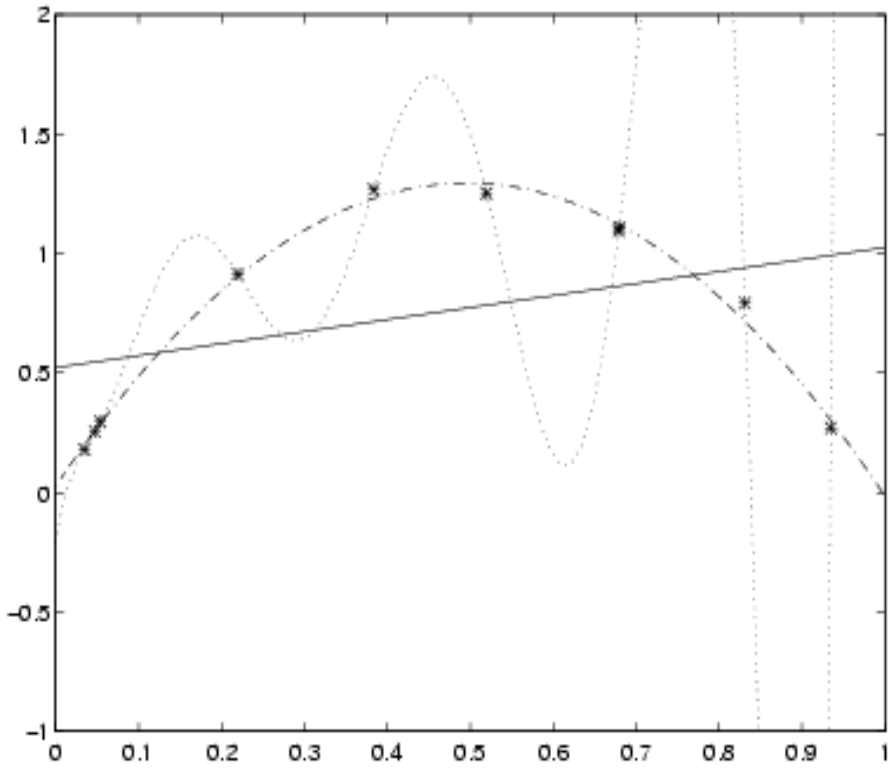


Figure 4: An example of minimum squared error polynomials of degrees 1, 2, and 9 for a set of 10 training points. The large degree polynomial demonstrates erratic behavior off the training set.

Approximation ratio:

$$\frac{\text{true error of selected hypothesis}}{\text{true error of best hypothesis considered}}$$

Cross validation (Ten-fold)

Structural risk minimization

$t = 20$	TRI	CVT	SRM	RIC	GCV	BIC	AIC	FPE	ADJ
25	1.00	1.06	1.14	7.54	5.47	15.2	22.2	25.8	1.02
50	1.06	1.17	1.39	224	118	394	585	590	1.12
75	1.17	1.42	3.62	5.8e3	3.9e3	9.8e3	1.2e4	1.2e4	1.24
95	1.44	6.75	56.1	6.1e5	3.7e5	7.8e5	9.2e5	8.2e5	1.54
100	2.41	1.1e4	2.2e4	1.5e8	6.5e7	1.5e8	1.5e8	8.2e7	3.02

$t = 30$	TRI	CVT	SRM	RIC	GCV	BIC	AIC	FPE	ADJ
25	1.00	1.08	1.17	4.69	1.51	5.41	5.45	2.72	1.06
50	1.08	1.17	1.54	34.8	9.19	39.6	40.8	19.1	1.14
75	1.19	1.37	9.68	258	91.3	266	266	159	1.25
95	1.45	6.11	419	4.7e3	2.7e3	4.8e3	5.1e3	4.0e3	1.51
100	2.18	643	1.6e7	1.6e7	1.6e7	1.6e7	1.6e7	1.6e7	2.10

Table 1: Fitting $f(x) = \text{step}(x \geq 0.5)$ with $P_x = U(0,1)$ and $\sigma = 0.05$. Tables give distribution of approximation ratios achieved at training sample size $t = 20$ and $t = 30$, showing percentiles of approximation ratios achieved in 1000 repeated trials.

$t = 20$	TRI	CVT	SRM	RIC	GCV	BIC	AIC	FPE	ADJ
25	2.04	1.03	1.00	1.00	1.06	1.00	1.01	1.58	1.02
50	3.11	1.37	1.33	1.34	1.94	1.35	1.61	18.2	1.32
75	3.87	2.23	2.30	2.13	10.0	2.75	4.14	1.2e3	1.83
95	5.11	9.45	8.84	8.26	5.0e3	11.8	82.9	1.8e5	3.94
100	8.92	105	526	105	2.0e7	2.1e3	2.7e5	2.4e7	6.30

$t = 30$	TRI	CVT	SRM	RIC	GCV	BIC	AIC	FPE	ADJ
25	1.50	1.00	1.00	1.00	1.00	1.00	1.00	1.02	1.01
50	3.51	1.16	1.03	1.05	1.11	1.02	1.08	1.45	1.27
75	4.15	1.64	1.45	1.48	2.02	1.39	1.88	6.44	1.60
95	5.51	5.21	5.06	4.21	26.4	5.01	19.9	295	3.02
100	9.75	124	1.4e3	20.0	9.1e3	28.4	9.4e3	1.0e4	8.35

Table 4: Fitting $f(x) = \sin^2(2\pi x)$ with $\mathrm{P}_x = U(0, 1)$ and $\sigma = 0.05$. Tables give distribution of approximation ratios achieved at training sample size $t = 20$ and $t = 30$, showing percentiles of approximation ratios achieved in 1000 repeated trials.

Bound on Error of TRI Relative to Best Hypothesis Considered

Proposition 1 *Let h_m be the optimal hypothesis in the sequence $h_0, h_1, \ldots$ (that is, $h_m = \arg\min_{h_k} d(h_k, \mathrm{P}_{Y|X})$) and let h_ℓ be the hypothesis selected by TRI. If (i) $m \leq \ell$ and (ii) $d(h_m, \widehat{\mathrm{P}_{Y|X}}) \leq d(h_m, \mathrm{P}_{Y|X})$ then*

$$d(h_\ell, \mathrm{P}_{Y|X}) \quad \leq \quad 3d(h_m, \mathrm{P}_{Y|X}) \tag{6}$$

Extension to TRI:
Adjust for expected bias of training data estimates

[Schuurmans & Southey, MLJ 2002]

Procedure ADJ

- Given hypothesis sequence $h_0, h_1, ...$
- For each hypothesis h_ℓ in the sequence
  - multiply its estimated distance to the target $d(h_\ell, \mathrm{P}_{Y|X})$ by the worst ratio of unlabeled and labeled distance to some predecessor h_k to obtain an adjusted distance estimate $d(h_\ell, \mathrm{P}_{Y|X}) = d(h_\ell, \mathrm{P}_{Y|X}) \frac{d(h_k, h_\ell)}{d(h_k, h_\ell)}$.
- Choose the hypothesis h_n with the smallest adjusted distance $d(h_n, \mathrm{P}_{Y|X})$.

Experimental results: averaged over multiple target functions, outperforms TRI