Announcements

- Homework 3 was due today.
- Homework 4 has been posted.
- Solutions to HW1 have been posted; more coming soon.
- Grader contact information is in the course syllabus, posted in "Overview" section on D2L.

Today's Lecture

- Feasibility of learning (part 1)
  - Understanding ML and its feasibility
  - Generalization error
  - Marbles and bins
  - Hoeffding inequality

FEASIBILITY OF LEARNING (part 1)

WHAT CAN MAKE A ML PROBLEM FEASIBLE?

IDEAS
- IF TRAINING DATA IS LINEARLY SEPARABLE
- IF WE CAN CONSTRUCT A SUITABLE OBJECTIVE FCN.
- IF WE MAKE ASSUMPTION(S) ON THE PROBLEM OR DATA.

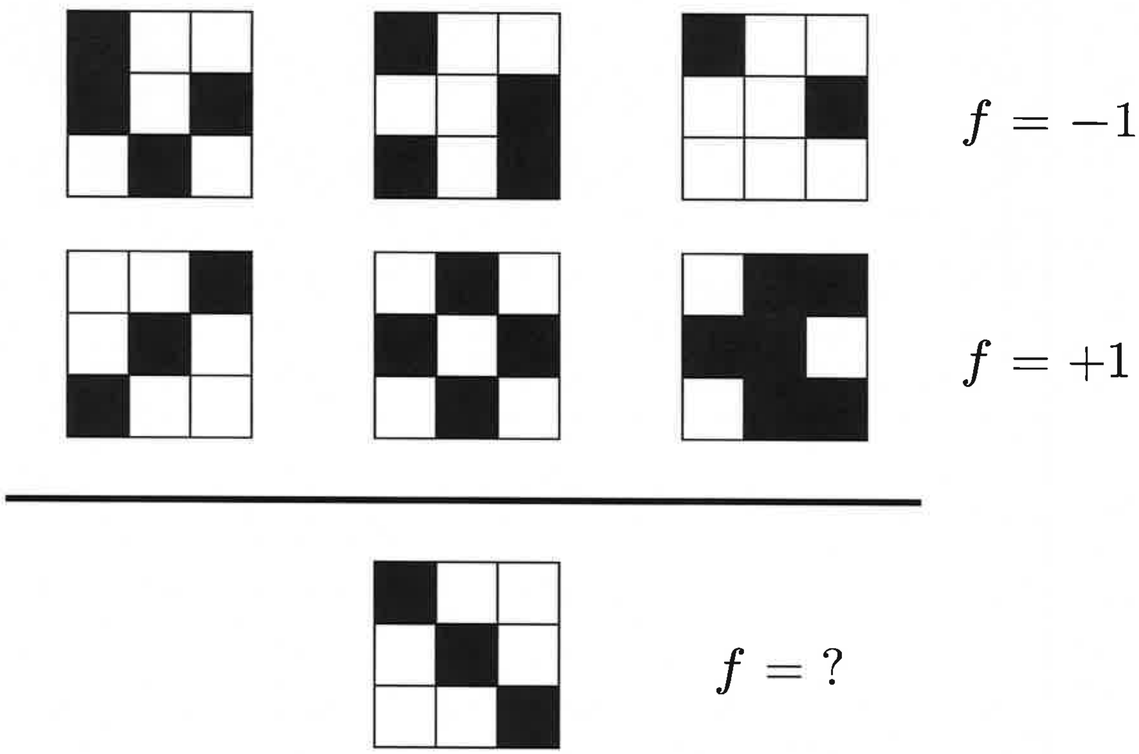
  e.g.: #BLACK SQUARES IS SIGNIFICANT.
  OR, $\exists$ SOME UNDERLYING CONCEPT THAT DEFINES CLASS LABELS.

CONCEPT LEARNING EXAMPLE:
- ASSUMPTION: $\exists$ A SIMPLE UNDERLYING CONCEPT THAT DEFINES THE CORRECT HYPOTHESIS. (e.g., POWERS OF 4).

POLYNOMIAL CURVE FIT EXAMPLE:

- LIMIT THE COMPLEXITY OF THE MODEL. [e.g.: BY d SMALL ENOUGH; OR USING PRIOR / REGULARIZER.

A Learning puzzle



SUMMARY:

TECHNIQUES THAT MIGHT ENABLE LEARNING

- USING A MODEL
- IMPOSING OUR OWN BIAS OR PRIOR KNOWLEDGE
- MAKING OTHER ASSUMPTIONS.

  (e.g., TYPE OF CONCEPT THAT IS LIKELY; OR DOMAIN OF PROBLEM (e.g., MATH OR ENGINEERING VS. ART FOR NUMBERS GAME).

IN ML TERMS, WE CAN DO THIS BY:

- CHOICE OF HYPOTHESIS SET $\mathcal{H}$.
- USE OF BAYESIAN PRIORS.
- DESIGN OF FEATURE SET, OR DESIGNING THE ALGORITHM THAT LEARNS THE SET OF FEATURES.

GENERALIZATION ERROR

LET $E_{in}(h)$ = IN-SAMPLE ERROR

(ERROR ON TRAINING DATA $\mathcal{D}_{Tr}$).

$$= \frac{1}{N} \sum_{n=1}^{N} [\![h(\underline{x}_n) \neq f(\underline{x}_n)]\!]$$

hypoth. (e.g., $\hat{f}(\underline{x}_n)$ or $\hat{y}(\underline{x}_n)$)

TRUE TARGET FCN. $= y_n$.

$E_{OUT}(h)$ = OUT-OF-SAMPLE ERROR

(ERROR ON UNKNOWNS)

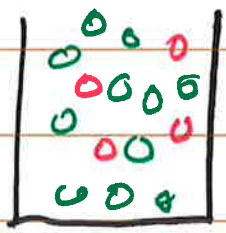
$$= P\left[h(\underline{x}) \neq f(\underline{x}) \right]$$

$E_{test}(h)$ = ERROR ON TEST SET $\mathcal{D}_{Test}$

$$= \frac{1}{N_{Test}} \sum_{\underline{x}_i \in \mathcal{D}_{Test}} [\![h(\underline{x}_i) \neq f(\underline{x}_i)]\!]$$

WE CAN MEASURE E_{in} AND E_{Test}; WE WANT TO KNOW E_{OUT}.

MARBLES AND BINS, μ AND ν



PICK A MARBLE

LET $\mu = P(\text{red})$

THEN $1-\mu = P(\text{green})$

μ IS UNKNOWN TO US.

DRAW N DATA POINTS INDEPENDENTLY (WITH REPLACEMENT) z_i, $i=1,2,\cdots,N$.

EST. $\hat{\mu} = \dfrac{\#\text{RED MARBLES}}{N} \triangleq \nu$.

HOW ACCURATE IS OUR ESTIMATE?

HOEFFDING INEQUALITY:

$$P\big[\,|\nu-\mu| > \epsilon\,\big] \leq 2e^{-2\epsilon^2 N}$$

FOR ANY $\epsilon > 0$

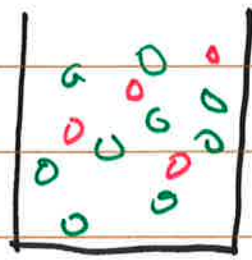
ϵ IS OUR "TOLERANCE"

ν IS THE ONLY RANDOM QUANTITY.

HOEFFDING INEQ. AND FEAS. OF LEARNING

CONSIDER: $C = 2$ CLASS PROBLEM.

FOR A HYPOTHESIS h (e.g., $h(x) = \text{sgn}\{\underline{w}^T \underline{x}\}$ WITH $\underline{w}$ SPECIFIED).



MARBLE/BIN SCENARIO	M.L. SCENARIO
ENTIRE SETS OF MARBLES IN BIN	SAMPLES FROM ALL ~~DATA PO~~ FEATURE SPACE $\mathcal{X}$, DRAWN ACCORDING TO $p(\underline{x})$.
COLOR OF MARBLE:	
GREEN	$h(\underline{x}_i) = f(\underline{x}_i)$ (CORRECT)
RED	$h(\underline{x}_i) \neq f(\underline{x}_i)$ (ERROR)
SET OF MARBLES DRAWN	(TRAINING) DATASET $\mathcal{D}$, ($\mathcal{D}_{Tr}$, $\mathcal{D}_{Test}$, $\mathcal{D}_{valid}$

ν = % OF MARBLES DRAWN THAT ARE RED $E_{in}(h)$, OR $E_{Test}(h)$, ...

$\mu = P[\text{marble is red}]$ $E_{out}(h)$

NOTE THAT THE MARBLE'S COLOR DEPENDS ON h.

HOEFFDING INEQ.:

$$P[|\mu-\nu|>\epsilon] \leq 2e^{-2\epsilon^2 N}$$

FOR ANY $\epsilon > 0$.

$$\boxed{P\big[\,|E_{out}(h) - E_{in}(h)| > \epsilon\,\big] \leq 2e^{-2\epsilon^2 N} \quad \text{FOR ANY } \epsilon > 0}$$

NOTE: DATASET $\mathcal{D}$ (TRAINING, TEST, ETC.), DRAWN AT RANDOM ACCORDING TO $p(\underline{x})$.

PROCEDURE FOR HOEFFDING INEQ. TO BE VALID:

1. SPECIFY h. ($\Rightarrow$, WITH $p(\underline{x})$, DETERMINES THE BIN OF MARBLES)
2. DRAW $\mathcal{D}$
3. GET BOUNDS ON $E_{out}(h)$, GIVEN $E_{\mathcal{D}}(h)$

ML PARADIGM $\Rightarrow$

1. COLLECT DATASET $\mathcal{D}$
2. CONSTRUCT HYPOTHESIS SET $\mathcal{H} = \{h_m\}_{m=1}^{M}$
3. TRAIN TO FIND h_g (best hypothesis).
4. CALCULATE $E_{in}(h_g)$ OR $E_{Test}(h_g)$.
5. WOULD REALLY WANT TO KNOW: $E_{OUT}(h_g)$.