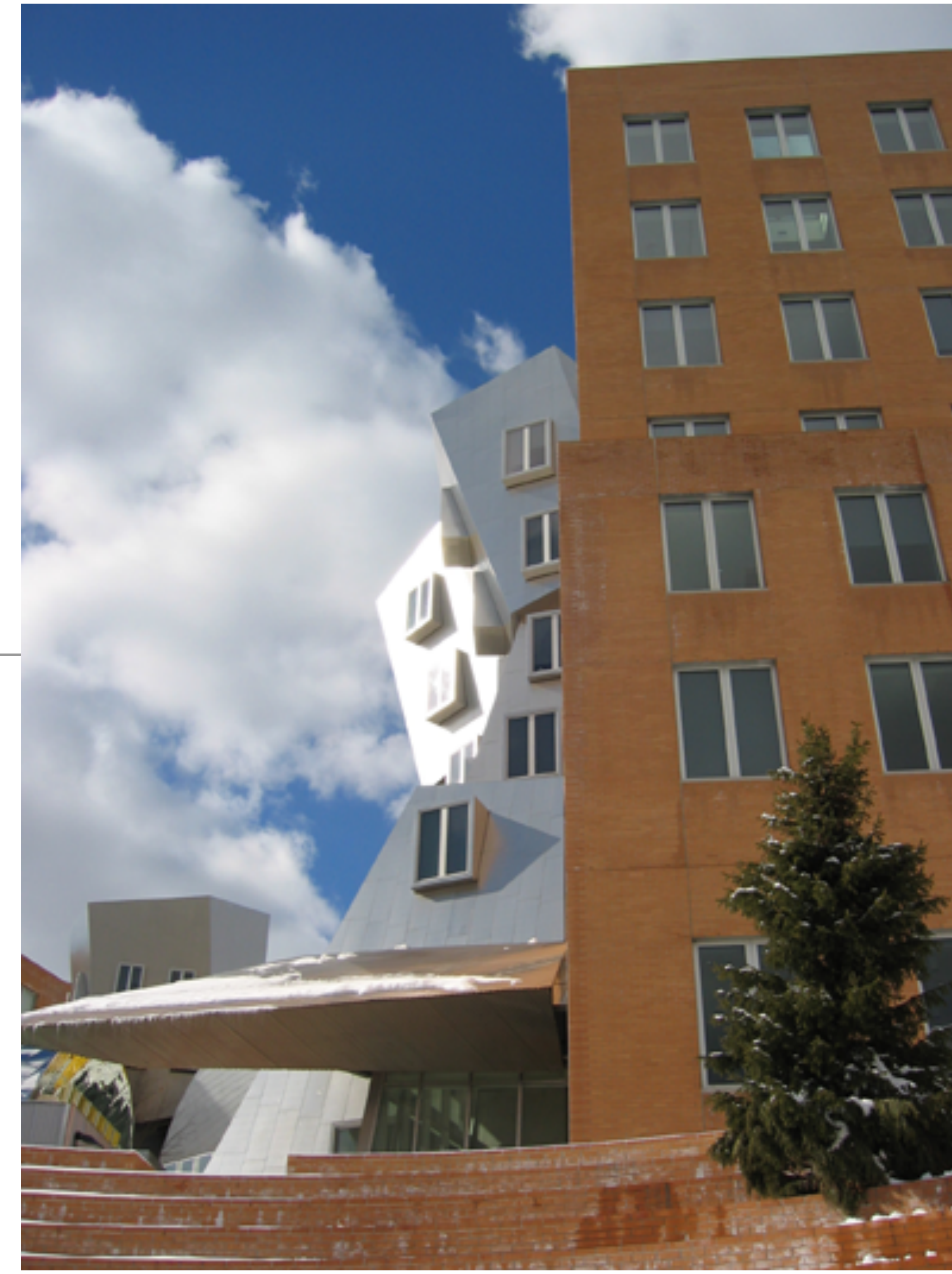
David's Lecture on Learning with Noisy Data



Massachusetts Institute of Technology

- If true distribution is $P(X, Y, \tilde{Y})$, with features X, true label Y, observed label $\tilde{Y}$, with $Y, \tilde{Y} \in \{+1, -1\}$
- Data are sampled from $P(X, \tilde{Y}) = \sum_y P(X, Y = y, \tilde{Y})$
- Assume that $P(X, Y, \tilde{Y}) = P(X, Y)P(\tilde{Y} | Y)$, i.e., $\tilde{Y} \perp X | Y$, or $P(\tilde{Y} | Y) = P(\tilde{Y} | Y, X)$
  - This may not hold, though even if not, in practice things work out.
- Suppose $P(\tilde{Y} = -1 | Y = 1) = \rho_+$, and $P(\tilde{Y} = 1 | Y = -1) = \rho_-$
- and assume $\rho_+ + \rho_- < 1$
- If we knew $\eta(x) = P(Y = 1 | X)$, then we could predict optimally
  - Bayes optimal classifier; if $\eta(X) > .5$, classify as $Y = 1$, else $Y = -1$
- We can approximate $\eta(X)$

- $\tilde{\eta}(X) = P(\tilde{Y} = 1 \mid X)$, but this is observable from data
- $\tilde{\eta}(X) = P(\tilde{Y} = 1, Y = 1 \mid X) + P(\tilde{Y} = 1, Y = -1 \mid X)$
- $\tilde{\eta}(X) = P(Y = 1 \mid X)P(\tilde{Y} = 1 \mid Y = 1) + P(Y = -1 \mid X)P(\tilde{Y} = 1 \mid Y = -1)$
- $= \eta(X)(1 - \rho_+) + (1 - \eta(X))\rho_-$
- $= \eta(X)(1 - \rho_+ - \rho_-) + \rho_-$
- If flip rates are 0, $\eta(X) = \tilde{\eta}(X)$
- Positive-only ?
-

Can we improve learning algorithm to work better with label noise?

- Methods from Natarajan 2013
  - 1. re-weight loss function
  - 2. modify (suitably symmetric) loss function
- Empirical risk minimization
  - $\min_f \frac{1}{n} \sum_{i=1}^{n} l(f(x_i), y_i), f : X \to \mathbb{R}$
  - loss functions
    - hinge: $L(t, y) = \max(0,1 - yt)$
    - 0-1: $l_{0/1}(t, y) = I[\text{sign}(t) \neq y]$
    - logistic: $l(t, y) = \lg(1 + e^{-yt})$
-

- Method 1
  - $l_\alpha(t, y) = (1 - \alpha)\mathbf{1}[y = 1]l(t,1) + \alpha\mathbf{1}[y = -1]l(t, -1)$
  - If $\alpha = 1/2$, $l_{\frac{1}{2}}(t, y) = \frac{1}{2}l(t, y)$
  - $\alpha* = \frac{1 - \rho_+ + \rho_-}{2}$
- Method 2
  - $\tilde{l}(t, y) = \frac{(1 - l_{-y})l(t, y) - \rho_y l(t, -y)}{1 - \rho_+ - \rho_-}$; l has to be symmetric
  - We want $\mathbb{E}_{\tilde{y}\sim y}[\tilde{l}(t, \tilde{y})] = l(t, y)$
  - for $y = 1$: $(1 - \rho_+)\tilde{l}(t,1) + \rho_+\tilde{l}(t, -1) = l(t,1)$
  - for $y = 0$: $(1 - \rho_-)\tilde{l}(t, -1) + \rho_-\tilde{l}(t,1) = l(t, -1)$
  - 2 equations in 2 unknowns, can be solved
- Problem of instance-dependent label noise (contrary to our assumption, $\tilde{Y} \perp X \mid Y$
  - Assume noise is largest near decision boundaries

Continuously predicted electronic phenotype

- Many “intermediate level” tags; may be able to extract these from notes.
- Commonly, derived manual rules
- Instead, use noisy labels within training data: anchors
  - E.g., insulin in meds, ICD code in discharge summary, …
  - Assume $A \perp X \,|\, Y$ where X are the other features; A is noisy label
    - May need to modify X to make this (more) true
    - Normally, we’re not able to predict from noisy label, but here the anchor has useful data.
  - Learn classifier to predict whether the anchor appears based on al the other features
  - At test time if anchor is present, predict 1; else…
- At BI, asked questions at time of discharge, to learn the actual Y variable; for some patients.