Who Needs Labels Anyway? 
or
Unsupervised Supervised Learning

Guy Lebanon

Georgia Institute of Technology
Perform supervised learning tasks, **without any labels**

- **Task 1:** Given a linear classifier estimate its expected hinge-loss or log-loss
- **Task 2:** Train a margin-based classifier (SVM, log-reg)
- **Task 3:** Given a classifier, estimate its expected error rate
- **Task 4:** Given a regression model, estimate its MSE

**Motivation:**

- Training without labels: labels completely unavailable, domain adaptation, assist semi-supervised and active learning
- Available classifier but no labeled data: privacy, domain adaptation, intellectual property
Perform supervised learning tasks, **without any labels**

- **Task 1:** Given a linear classifier **estimate** its expected hinge-loss or log-loss
- **Task 2:** **Train** a margin-based classifier (SVM, log-reg)
  - **Task 3:** Given a classifier, **estimate** its expected error rate
  - **Task 4:** Given a regression model, **estimate** its MSE

**Motivation:**

- Training without labels: labels completely unavailable, domain adaptation, assist semi-supervised and active learning
- Available classifier but no labeled data: privacy, domain adaptation, intellectual property
Unsupervised Supervised Learning

Perform supervised learning tasks, without any labels

- Task 1: Given a linear classifier estimate its expected hinge-loss or log-loss
- Task 2: Train a margin-based classifier (SVM, log-reg)
- Task 3: Given a classifier, estimate its expected error rate
- Task 4: Given a regression model, estimate its MSE

Motivation:

- Training without labels: labels completely unavailable, domain adaptation, assist semi-supervised and active learning
- Available classifier but no labeled data: privacy, domain adaptation, intellectual property
Perform supervised learning tasks, without any labels

- Task 1: Given a linear classifier estimate its expected hinge-loss or log-loss
- Task 2: Train a margin-based classifier (SVM, log-reg)
- Task 3: Given a classifier, estimate its expected error rate
- Task 4: Given a regression model, estimate its MSE

Motivation:

- Training without labels: labels completely unavailable, domain adaptation, assist semi-supervised and active learning
- Available classifier but no labeled data: privacy, domain adaptation, intellectual property
Perform supervised learning tasks, **without any labels**

- **Task 1**: Given a linear classifier **estimate** its expected hinge-loss or log-loss
- **Task 2**: **Train** a margin-based classifier (SVM, log-reg)
- **Task 3**: Given a classifier, **estimate** its expected error rate
- **Task 4**: Given a regression model, **estimate** its MSE

**Motivation:**

- Training without labels: labels completely unavailable, domain adaptation, assist semi-supervised and active learning
- Available classifier but no labeled data: privacy, domain adaptation, intellectual property
Perform supervised learning tasks, without any labels

- Task 1: Given a linear classifier, estimate its expected hinge-loss or log-loss
- Task 2: Train a margin-based classifier (SVM, log-reg)
- Task 3: Given a classifier, estimate its expected error rate
- Task 4: Given a regression model, estimate its MSE

Motivation:

- Training without labels: labels completely unavailable, domain adaptation, assist semi-supervised and active learning
- Available classifier but no labeled data: privacy, domain adaptation, intellectual property
Task 1: Estimate Expected Margin-Based Loss

- \( \hat{Y} = \text{sign} \sum_{j=1}^{d} \theta_j X_j \) with \( Y \in \{-1,+1\}, \quad X \in \mathbb{R}^d \)

- Expected loss \( R(\theta) = \mathbb{E}_{p(X,Y)} L_i(Y, f_\theta(X)) \),

\[
\begin{align*}
    f_\theta(X) & \overset{\text{def}}{=} \sum_{j=1}^{d} \theta_j X_j \\
    L_1(Y, f_\theta(X)) & = \exp(-Yf_\theta(X)) \quad \text{exp loss} \\
    L_2(Y, f_\theta(X)) & = \log(1 + \exp(-Yf_\theta(X))) \quad \text{log-loss} \\
    L_3(Y, f_\theta(X)) & = (1 - Yf_\theta(X))^+ \quad \text{hinge-loss}
\end{align*}
\]

plays key role in SVM, logistic regression, boosting

- Estimate \( R(\theta) \) using labeled data

\[
R_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} L(Y^{(i)}, f_\theta(X^{(i)}))
\]

- Estimate \( R(\theta) \) using unlabeled data \( \hat{R}_n(\theta; X^{(1)}, \ldots, X^{(n)}) \)

\[
\hat{R}_n(\theta; X^{(1)}, \ldots, X^{(n)}) \to R(\theta) \quad \text{with probability 1}
\]
Task 1: Estimate Expected Margin-Based Loss

\[ \hat{Y} = \text{sign} \sum_{j=1}^{d} \theta_j X_j \text{ with } Y \in \{-1, +1\}, \quad X \in \mathbb{R}^d \]

- Expected loss \( R(\theta) = \mathbb{E}_{p(X,Y)} L_i(Y, f_\theta(X)) \),
  \[ f_\theta(X) \overset{\text{def}}{=} \sum_{j=1}^{d} \theta_j X_j \]
  \[ L_1(Y, f_\theta(X)) = \exp (-Yf_\theta(X)) \quad \text{exp loss} \]
  \[ L_2(Y, f_\theta(X)) = \log (1 + \exp (-Yf_\theta(X))) \quad \text{log-loss} \]
  \[ L_3(Y, f_\theta(X)) = (1 - Yf_\theta(X))_+ \quad \text{hinge-loss} \]

plays key role in SVM, logistic regression, boosting

- Estimate \( R(\theta) \) using labeled data
  \[ R_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} L(Y^{(i)}, f_\theta(X^{(i)})) \]

- Estimate \( R(\theta) \) using unlabeled data \( \hat{R}_n(\theta; X^{(1)}, \ldots, X^{(n)}) \)
  \[ \hat{R}_n(\theta; X^{(1)}, \ldots, X^{(n)}) \rightarrow R(\theta) \quad \text{with probability 1} \]
Task 1: Estimate Expected Margin-Based Loss

\( \hat{Y} = \text{sign} \sum_{j=1}^{d} \theta_j X_j \) with \( Y \in \{-1, +1\}, \quad X \in \mathbb{R}^d \)

- Expected loss \( R(\theta) = \mathbb{E}_{p(X,Y)} L_i(Y, f_\theta(X)) \),
  \( f_\theta(X) \overset{\text{def}}{=} \sum_{j=1}^{d} \theta_j X_j \)

  \[
  L_1(Y, f_\theta(X)) = \exp(-Yf_\theta(X)) \quad \text{exp loss}
  \]

  \[
  L_2(Y, f_\theta(X)) = \log(1 + \exp(-Yf_\theta(X))) \quad \text{log-loss}
  \]

  \[
  L_3(Y, f_\theta(X)) = (1 - Yf_\theta(X))_+ \quad \text{hinge-loss}
  \]

plays key role in SVM, logistic regression, boosting

- Estimate \( R(\theta) \) using labeled data
  \( R_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} L(Y^{(i)}, f_\theta(X^{(i)})) \)

- Estimate \( R(\theta) \) using unlabeled data \( \hat{R}_n(\theta; X^{(1)}, \ldots, X^{(n)}) \)

  \[ \hat{R}_n(\theta; X^{(1)}, \ldots, X^{(n)}) \to R(\theta) \quad \text{with probability 1} \]
Task 1: Estimate Expected Margin-Based Loss

\[ \hat{Y} = \text{sign} \sum_{j=1}^{d} \theta_j X_j \text{ with } Y \in \{-1, +1\}, \quad X \in \mathbb{R}^d \]

- Expected loss \( R(\theta) = \mathbb{E}_{p(X,Y)} L_i(Y, f_\theta(X)) \),
  \[ f_\theta(X) \overset{\text{def}}{=} \sum_{j=1}^{d} \theta_j X_j \]
  \[ L_1(Y, f_\theta(X)) = \exp (-Y f_\theta(X)) \quad \text{exp loss} \]
  \[ L_2(Y, f_\theta(X)) = \log (1 + \exp (-Y f_\theta(X))) \quad \text{log-loss} \]
  \[ L_3(Y, f_\theta(X)) = (1 - Y f_\theta(X))_+ \quad \text{hinge-loss} \]

plays key role in SVM, logistic regression, boosting

- Estimate \( R(\theta) \) using labeled data
  \[ R_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} L(Y^{(i)}, f_\theta(X^{(i)})) \]

- Estimate \( R(\theta) \) using unlabeled data \( \hat{R}_n(\theta; X^{(1)}, \ldots, X^{(n)}) \)
  \[ \hat{R}_n(\theta; X^{(1)}, \ldots, X^{(n)}) \to R(\theta) \quad \text{with probability 1} \]
Assumptions

Two important assumptions that hold in many practical cases

- Label marginals $p(y = 1), p(y = -1)$ are unequal and known
  - Example: medical diagnosis (disease frequency)
  - Example: medical prognosis (life expectancy)
  - Example: OCR (letter frequency in English)
  - Example: Object recognition (object frequency in scene)

- $f_\theta(X)|Y = \sum_{i=1}^{d} \theta_i X_i|Y$ is normally distributed
  - Motivated by central limit theorems for high $d$
  - Holds in practice for high $d$
Assumptions

Two important assumptions that hold in many practical cases

- Label marginals $p(y = 1), p(y = -1)$ are unequal and known
  - Example: medical diagnosis (disease frequency)
  - Example: medical prognosis (life expectancy)
  - Example: OCR (letter frequency in English)
  - Example: Object recognition (object frequency in scene)

- $f_\theta(X)|Y = \sum_{i=1}^{d} \theta_i X_i|Y$ is normally distributed
  - Motivated by central limit theorems for high $d$
  - Holds in practice for high $d$
Assumptions

Two important assumptions that hold in many practical cases

- Label marginals $p(y = 1), p(y = -1)$ are unequal and known
  - Example: medical diagnosis (disease frequency)
  - Example: medical prognosis (life expectancy)
  - Example: OCR (letter frequency in English)
  - Example: Object recognition (object frequency in scene)

- $f_{\theta}(X)\mid Y = \sum_{i=1}^{d} \theta_i X_i \mid Y$ is normally distributed
  - Motivated by central limit theorems for high $d$
  - Holds in practice for high $d$
Two important assumptions that hold in many practical cases

- Label marginals $p(y = 1), p(y = -1)$ are unequal and known
  - Example: medical diagnosis (disease frequency)
  - Example: medical prognosis (life expectancy)
  - Example: OCR (letter frequency in English)
  - Example: Object recognition (object frequency in scene)

- $f_\theta(X) | Y = \sum_{i=1}^{d} \theta_i X_i | Y$ is normally distributed
  - Motivated by central limit theorems for high $d$
  - Holds in practice for high $d$
Two important assumptions that hold in many practical cases

- Label marginals $p(y = 1), p(y = -1)$ are unequal and known
  - Example: medical diagnosis (disease frequency)
  - Example: medical prognosis (life expectancy)
  - Example: OCR (letter frequency in English)
  - Example: Object recognition (object frequency in scene)

- $f_{\theta}(X)|Y = \sum_{i=1}^{d} \theta_i X_i|Y$ is normally distributed
  - Motivated by central limit theorems for high $d$
  - Holds in practice for high $d$
Assumptions

Two important assumptions that hold in many practical cases

- Label marginals $p(y = 1), p(y = -1)$ are unequal and known
  - Example: medical diagnosis (disease frequency)
  - Example: medical prognosis (life expectancy)
  - Example: OCR (letter frequency in English)
  - Example: Object recognition (object frequency in scene)

- $f_\theta(X) | Y = \sum_{i=1}^{d} \theta_i X_i | Y$ is normally distributed
  - Motivated by central limit theorems for high $d$
  - Holds in practice for high $d$
Two important assumptions that hold in many practical cases

- Label marginals $p(y = 1), p(y = -1)$ are unequal and known
  - Example: medical diagnosis (disease frequency)
  - Example: medical prognosis (life expectancy)
  - Example: OCR (letter frequency in English)
  - Example: Object recognition (object frequency in scene)

- $f_\theta(X)|Y = \sum_{i=1}^{d} \theta_i X_i|Y$ is normally distributed
  - Motivated by central limit theorems for high $d$
  - Holds in practice for high $d$
Two important assumptions that hold in many practical cases

- Label marginals $p(y = 1), p(y = -1)$ are unequal and known
  - Example: medical diagnosis (disease frequency)
  - Example: medical prognosis (life expectancy)
  - Example: OCR (letter frequency in English)
  - Example: Object recognition (object frequency in scene)

- $f_\theta(X)|Y = \sum_{i=1}^{d} \theta_i X_i|Y$ is normally distributed
  - Motivated by central limit theorems for high $d$
  - Holds in practice for high $d$
Proposition (de-Moivre)

If $Z_i, i \in \mathbb{N}$ are iid with expectation $\mu$ and variance $\sigma^2$ and 
$\bar{Z}_d = d^{-1} \sum_{i=1}^{d} Z_i$ then we have the following convergence in distribution

$$\sqrt{d}(\bar{Z}_d - \mu)/\sigma \xrightarrow{d \to \infty} N(0, 1)$$

A sum of many iid RVs is normally distributed
Proposition (Lindberg)

For $Z_i, i \in \mathbb{N}$ independent with expectation $\mu_i$ and variance $\sigma_i^2$, and denoting $s_d^2 = \sum_{i=1}^{d} \sigma_i^2$, we have the following convergence in distribution as $d \to \infty$

$$s_d^{-1} \sum_{i=1}^{d} (Z_i - \mu_i) \rightsquigarrow N(0, 1)$$

if the following condition holds for every $\epsilon > 0$

$$\lim_{d \to \infty} s_d^{-2} \sum_{i=1}^{d} \mathbb{E}(Z_i - \mu_i)^2 1\{|X_i - \mu_i| > \epsilon s_d\} = 0. \quad (1)$$

A sum of many independent RVs is normally distributed (if the variance of the sum is not dominated by a few variables)
Definition

The random variables $Z_i, i \in \mathbb{N}$ are said to be $m(k)$-dependent if whenever $s - r > m(k)$ the two sets \{\(Z_1, \ldots, Z_r\), \(Z_s, \ldots, Z_k\)\} are independent.

Proposition (Berk)

For each $k$ let $d(k)$ and $m(k)$ be increasing sequences and suppose that $Z_1^{(k)}, \ldots, Z_d(k)$ is an $m(k)$-dependent RV sequences. If

1. $\mathbb{E}|Z_i^{(k)}|^2 \leq M$ for all $i$ and $k$
2. $\text{Var}(Z_{i+1}^{(k)} + \ldots + Z_j^{(k)}) \leq (j - i)K$ for all $i, j, k$
3. $\lim_{k \to \infty} \text{Var}(Z_1^{(k)} + \ldots + Z_{d(k)}^{(k)})/d(k)$ exists and is non-zero
4. $\lim_{k \to \infty} m^2(k)/d(k) = 0$

then \(\sum_{i=1}^{d(k)} \frac{Z_i^{(k)}}{\sqrt{d(k)}}\) is asymptotically normal as $k \to \infty$. 
Definition

A graph $G = (\mathcal{V}, \mathcal{E})$ indexing r.v. is called a dependency graph if for any pair of disjoint subsets of $\mathcal{V}$, $A_1$ and $A_2$ such that no edge in $\mathcal{E}$ has one endpoint in $A_1$ and the other in $A_2$, we have independence between $\{Z_i : i \in A_1\}$ and $\{Z_i : i \in A_2\}$.

Proposition (Rinott)

Let $Z_1, \ldots, Z_n$ be random variables having a dependency graph with max degree $\max_{v \in \mathcal{V}} d(v) < D$, and satisfying $|Z_i - E(Z_i)| \leq B$ a.s., $\forall i$, $E(\sum_{i=1}^{n} Z_i) = \lambda$ and $\text{Var}(\sum_{i=1}^{n} Z_i) = \sigma^2 > 0$

\[
\sup_{w \in \mathbb{R}} \left| P \left( \frac{\sum_{i=1}^{n} Z_i - \lambda}{\sigma} \leq w \right) - \Phi(w) \right| \leq \frac{1}{\sigma} \left( \frac{1}{\sqrt{2\pi}} DB + 16 \left( \frac{n}{\sigma^2} \right)^{1/2} D^{3/2} B^2 + 10 \left( \frac{n}{\sigma^2} \right) D^2 B^3 \right)
\]

where $\Phi$ is the CDF corresponding to a $N(0,1)$ distribution.
RCV1 text data

face images

random $\theta$

Fisher's LDA

log. reg.

MNIST digit images
RCV1 text data

face images

MNIST digit images

log. reg. ($l_1$ reg)

log. reg. ($l_2$ reg)
Idea 1: Rewrite expected risk using expectation over $Y$, $\alpha = f_\theta(X)$

$$R(\theta) = \mathbb{E}_{p(f_\theta(X), Y)} L(Y, f_\theta(X))$$

$$= p(y = 1) \int_{\mathbb{R}} p(f_\theta(X) = \alpha | y = 1) L(1, \alpha) \, d\alpha$$

$$+ p(y = -1) \int_{\mathbb{R}} p(f_\theta(X) = \alpha | y = -1) L(-1, \alpha) \, d\alpha.$$

where

$$f_\theta(X) | y = 1 \sim N(\mu_1, \sigma_1^2)$$

$$f_\theta(X) | y = -1 \sim N(\mu_{-1}, \sigma_{-1}^2)$$

$\mu_1, \sigma_1^2$ unknown parameters

$\mu_{-1}, \sigma_{-1}^2$ unknown parameters
Idea 2: Estimate \( \mu = (\mu_1, \mu_{-1}) \) and \( \sigma = (\sigma_1, \sigma_{-1}) \) by maximizing the (observed) likelihood of unlabeled data

\[
(\hat{\mu}^{(n)}, \hat{\sigma}^{(n)}) = \arg \max_{\mu, \sigma} \ell_n(\mu, \sigma) \quad \text{where}
\]

\[
\ell_n(\mu, \sigma) = \sum_{i=1}^{n} \log p(f_\theta(X^{(i)}))
\]

\[
= \sum_{i=1}^{n} \log \sum_{y^{(i)}} p(f_\theta(X^{(i)}), y^{(i)})
\]

\[
= \sum_{i=1}^{n} \log \sum_{y^{(i)}} p(y^{(i)}) p_{\mu y^{(i)}, \sigma y^{(i)}}(f_\theta(X^{(i)})|y^{(i)})
\]
Consistency

Under the two assumptions above, the plug-in estimator of the risk is consistent i.e., as we get more unlabeled data the estimator converges to the true risk (no labels)

\[
P \left( \lim_{n \to \infty} \hat{R}_n(\theta ; X^{(1)}, \ldots, X^{(n)}) = R(\theta) \right) = 1 \quad \text{where}
\]

\[
\hat{R}_n(\theta ; X^{(1)}, \ldots, X^{(n)}) = \sum_{y} p(y) \int_{\mathbb{R}} p_{\hat{\mu}_y^{(n)}, \hat{\sigma}_y^{(n)}}(f_{\theta}(X) = \alpha | y) L(y, \alpha) \, d\alpha.
\]

proof based on re-expressing the margin-based risk, consistency of the MLE in mixture models (up to permutations) and continuity arguments.
## Domain adaptation on 20-newsgroups data

| Data            | $R_n$  | $|R_n - \hat{R}_n|$ | $|R_n - \hat{R}_n|/R_n$ | $n$   | $p(Y = 1)$ |
|-----------------|--------|---------------------|-------------------------|-------|------------|
| sci vs. comp    | 0.7088 | 0.0093              | 0.013                   | 3590  | 0.8257     |
| sci vs. rec     | 0.641  | 0.0141              | 0.022                   | 3958  | 0.7484     |
| talk vs. rec    | 0.5933 | 0.0159              | 0.026                   | 3476  | 0.7126     |
| talk vs. comp   | 0.4678 | 0.0119              | 0.025                   | 3459  | 0.7161     |
| talk vs. sci    | 0.5442 | 0.0241              | 0.044                   | 3464  | 0.7151     |
| comp vs. rec    | 0.4851 | 0.0049              | 0.010                   | 4927  | 0.7972     |
Relative error $|R_n - \hat{R}_n|/R_n$ decreases with train set size, classifier accuracy, class imbalance; robust to misspecification of $p(y)$
Task 2: Train Margin-Based Classifiers

- We showed we can accurately estimate hinge-loss or log-loss $\hat{R}_n(\theta)$ for a given classifier without labeled data.

- $\Rightarrow$ Train a margin-based classifier (SVM, log-reg) by minimizing $\hat{R}_n(\theta)$ iteratively (without labels).

- Two algorithms for unsupervised empirical risk minimization:
  - Finite difference gradient descent
  - Grid search
Task 2: Train Margin-Based Classifiers

- We showed we can accurately estimate hinge-loss or log-loss $\hat{R}_n(\theta)$ for a given classifier without labeled data.
- $\Rightarrow$ Train a margin-based classifier (SVM, log-reg) by minimizing $\hat{R}_n(\theta)$ iteratively (without labels).

- Two algorithms for unsupervised empirical risk minimization:
  - Finite difference gradient descent
  - Grid search
We showed we can accurately estimate hinge-loss or log-loss $\hat{R}_n(\theta)$ for a given classifier without labeled data.

$\Rightarrow$ Train a margin-based classifier (SVM, log-reg) by minimizing $\hat{R}_n(\theta)$ iteratively (without labels).

Two algorithms for unsupervised empirical risk minimization:

- Finite difference gradient descent
- Grid search
Task 2: Train Margin-Based Classifiers

- We showed we can accurately estimate hinge-loss or log-loss \( \hat{R}_n(\theta) \) for a given classifier without labeled data.
- **⇒** Train a margin-based classifier (SVM, log-reg) by minimizing \( \hat{R}_n(\theta) \) iteratively (without labels).
- Two algorithms for unsupervised empirical risk minimization:
  - Finite difference gradient descent
  - Grid search
Task 2: Train Margin-Based Classifiers

- We showed we can accurately estimate hinge-loss or log-loss $\hat{R}_n(\theta)$ for a given classifier without labeled data.
- $\Rightarrow$ Train a margin-based classifier (SVM, log-reg) by minimizing $\hat{R}_n(\theta)$ iteratively (without labels).
- Two algorithms for unsupervised empirical risk minimization:
  - Finite difference gradient descent
  - Grid search
Under further assumptions on data and parameter spaces, \( \arg \min_{\theta} \hat{R}_n(\theta) \) converges to the expected risk minimizer

\[
P \left( \lim_{n \to \infty} \arg \min_{\theta \in \Theta} R_n(\theta) = \arg \min_{\theta \in \Theta} R(\theta) \right) = 1.
\]

Implication: In cases where \( \arg \min_{\theta} R(\theta) \) is the Bayes classifier we can retrieve the optimal classifier without a single labeled data point. Proof based on uniform convergence arguments.
Gradient descent

**Input:** $X^{(1)}, \ldots, X^{(n)} \in \mathbb{R}^d$, $p(Y)$, step size $\alpha$

Initialize $t = 0$, $\theta^{(t)} = \theta^0 \in \mathbb{R}^d$

**repeat**

Compute $f_{\theta^{(t)}}(X^{(j)}) = \langle \theta^{(t)}, X^{(j)} \rangle \ \forall j = 1, \ldots, n$

Estimate $(\hat{\mu}_1, \hat{\mu}_{-1}, \hat{\sigma}_1, \hat{\sigma}_{-1})$ by maximizing the likelihood

**for** $i = 1$ **to** $d$ **do**

Plug-in the estimates to approximate the gradient

$$\frac{\partial \hat{R}_n(\theta^{(t)})}{\partial \theta_i} = \frac{\hat{R}_n(\theta^{(t)} + h_i e_i) - \hat{R}_n(\theta^{(t)} - h_i e_i)}{2h_i}$$

**end for**

Form $\nabla \hat{R}_n(\theta^{(t)}) = (\frac{\partial \hat{R}_n(\theta^{(t)})}{\partial \theta_1^{(t)}}, \ldots, \frac{\partial \hat{R}_n(\theta^{(t)})}{\partial \theta_d^{(t)}})$

**Update** $\theta^{(t+1)} = \theta^{(t)} - \alpha \nabla \hat{R}_n(\theta^{(t)})$, $t = t + 1$

**until** convergence

**Output:** linear classifier $\theta_{\text{final}} = \theta^{(t)}$
Grid search

**Input:** $X^{(1)}, \ldots, X^{(n)} \in \mathbb{R}^d$, $p(Y)$, grid-size $\tau$

Initialize $\theta_i \sim \text{Uniform}(-2, 2)$ for all $i$

repeat
  for $i = 1$ to $d$ do
    Construct $\tau$ points grid in the range $[\theta_i - 4\tau, \theta_i + 4\tau]$
    Compute the risk estimate with all dimensions of $\theta^{(t)}$ fixed except for $[\theta^{(t)}]_i$ which is evaluated at each grid point.
    Set $[\theta^{(t+1)}]_i$ to the grid value that minimized the risk estimate
  end for
until convergence

**Output:** linear classifier $\theta^{\text{final}} = \theta$
USL classification, RCV1 text data

Grad descent and grid search (USL and SL); SL logistic regression
test error 0.07 ($n = 140000$, $d = 500$, $p(Y = 1) = 0.3$)
Grad descent and grid search (USL and SL); SL logistic regression test error 0.05 ($n = 10000$, $d = 784$, $p(Y = 1) = 0.53$)
Task 3: Estimate Error-Rate of a Given Classifier

- Given a classifier \( f \) estimate its 0-1 risk
  \[ R(f) = P(f \text{ predicts wrong class}) \]
  using unlabeled data

- Assumptions:
  - \( p(Y = 1), p(Y = -1) \) are known and unequal
  - classifier is probabilistic
  - classifier is more accurate than random guess
  - false positive rate equals false negative rate*.

Applies to any probabilistic classifier (not only margin-based),
works in both low and high dimensions, doesn’t generalize well to
training classifiers
Task 3: Estimate Error-Rate of a Given Classifier

- Given a classifier $f$ estimate its 0-1 risk
  \[ R(f) = P(f \text{ predicts wrong class}) \] using unlabeled data

- Assumptions:
  - $p(Y = 1), p(Y = -1)$ are known and unequal
  - classifier is probabilistic
  - classifier is more accurate than random guess
  - false positive rate equals false negative rate*

Applies to any probabilistic classifier (not only margin-based), works in both low and high dimensions, doesn’t generalize well to training classifiers
Given a classifier $f$ estimate its 0-1 risk

$$R(f) = P(f \text{ predicts wrong class})$$

using unlabeled data

Assumptions:

- $p(Y = 1), p(Y = -1)$ are known and unequal
- classifier is probabilistic
- classifier is more accurate than random guess
- false positive rate equals false negative rate

Applies to any probabilistic classifier (not only margin-based), works in both low and high dimensions, doesn’t generalize well to training classifiers
Given a classifier $f$ estimate its 0-1 risk
$$R(f) = P(f \text{ predicts wrong class})$$ using unlabeled data

**Assumptions:**
- $p(Y = 1), p(Y = -1)$ are known and unequal
- Classifier is probabilistic
- Classifier is more accurate than random guess
- False positive rate equals false negative rate

Applies to any probabilistic classifier (not only margin-based), works in both low and high dimensions, doesn’t generalize well to training classifiers.
Task 3: Estimate Error-Rate of a Given Classifier

- Given a classifier \( f \) estimate its 0-1 risk \( R(f) = P(f \text{ predicts wrong class}) \) using unlabeled data
- Assumptions:
  - \( p(Y = 1), p(Y = -1) \) are known and unequal
  - classifier is probabilistic
  - classifier is more accurate than random guess
  - false positive rate equals false negative rate*

Applies to any probabilistic classifier (not only margin-based), works in both low and high dimensions, doesn’t generalize well to training classifiers
Task 3: Estimate Error-Rate of a Given Classifier

- Given a classifier $f$ estimate its 0-1 risk
  \[ R(f) = P(f \text{ predicts wrong class}) \] using unlabeled data

- Assumptions:
  - $p(Y = 1), p(Y = -1)$ are known and unequal
  - classifier is probabilistic
  - classifier is more accurate than random guess
  - false positive rate equals false negative rate*.

Applies to any probabilistic classifier (not only margin-based), works in both low and high dimensions, doesn’t generalize well to training classifiers
Task 3: Estimate Error-Rate of a Given Classifier

Key 1: 0-1 risk is a function of $p(\hat{y}|y)$: $R(f) = g(p(\hat{y}|y))$

Key 2: Estimate $\hat{p}(\hat{y}|y)$ by maximizing the likelihood of the unlabeled data

$$\hat{p}(\cdot|\cdot) = \arg \max_{p(\cdot|\cdot)} \sum_{i=1}^{n} \log \sum_{y(i)} p(\hat{y}^{(i)}|y^{(i)}) p(y^{(i)})$$

and use the plug-in rule to estimate the risk

$$\hat{R}(f ; \hat{y}^{(1)}, \ldots, \hat{y}^{(n)}) = g(\hat{p}(\cdot|\cdot))$$
Task 3: Estimate Error-Rate of a Given Classifier

- Key 1: 0-1 risk is a function of $p(\hat{y}|y)$: $R(f) = g(p(\hat{y}|y))$
- Key 2: Estimate $\hat{p}(\hat{y}|y)$ by maximizing the likelihood of the unlabeled data

$$\hat{p}(\cdot|\cdot) = \arg\max_{p(\cdot|\cdot)} \sum_{i=1}^{n} \log \sum_{y(i)} p(\hat{y}^{(i)}|y^{(i)})p(y^{(i)})$$

and use the plug-in rule to estimate the risk

$$\hat{R}(f; \hat{y}^{(1)}, \ldots, \hat{y}^{(n)}) = g(\hat{p}(\cdot|\cdot))$$
Above scheme provides a (unlabeled) consistent estimator of the 0-1 risk: \( \lim \hat{R}_n(f ; X^{(1)}, \ldots, X^{(n)}) = R(f) \) w.p. 1.

Proof based on showing identifiability of \( p(y) \) in terms of the parameter \( p(\hat{y}|y) \) and invoking strong consistency of MLE.

Asymptotic variance may be derived and shows the risk estimation accuracy increases with class imbalance and classifier accuracy:

\[
J(\theta) = \frac{\alpha(2\alpha - 1)^2}{(\theta(2\alpha - 1) - \alpha + 1)^2} - \frac{(2\alpha - 1)^2(\alpha - 1)}{(\alpha - \theta(2\alpha - 1))^2}
\]

where \( \alpha = p(Y = 1) \) and \( \theta = 1 - R(f) \).
Above scheme provides a (unlabeled) consistent estimator of the 0-1 risk: \( \lim \hat{R}_n(f; X^{(1)}, \ldots, X^{(n)}) = R(f) \) w.p. 1.

Proof based on showing identifiability of \( p(y) \) in terms of the parameter \( p(\hat{y}|y) \) and invoking strong consistency of MLE.

Asymptotic variance may be derived and shows the risk estimation accuracy increases with class imbalance and classifier accuracy.

\[
J(\theta) = \frac{\alpha (2\alpha - 1)^2}{(\theta(2\alpha - 1) - \alpha + 1)^2} - \frac{(2\alpha - 1)^2(\alpha - 1)}{(\alpha - \theta(2\alpha - 1))^2}
\]

where \( \alpha = p(Y = 1) \) and \( \theta = 1 - R(f) \).
Above scheme provides a (unlabeled) consistent estimator of the 0-1 risk: \( \lim \hat{R}_n(f; X^{(1)}, \ldots, X^{(n)}) = R(f) \) w.p. 1.

Proof based on showing identifiability of \( p(y) \) in terms of the parameter \( p(\hat{y}|y) \) and invoking strong consistency of MLE

Asymptotic variance may be derived and shows the risk estimation accuracy increases with class imbalance and classifier accuracy

\[
J(\theta) = \frac{\alpha(2\alpha - 1)^2}{(\theta(2\alpha - 1) - \alpha + 1)^2} - \frac{(2\alpha - 1)^2(\alpha - 1)}{(\alpha - \theta(2\alpha - 1))^2}
\]

where \( \alpha = p(Y = 1) \) and \( \theta = 1 - R(f) \).
Maximizing joint likelihood for $p_j(\hat{y}|y)$ for multiple classifiers $f_1, \ldots, f_k$ provides collaborative estimation and improved accuracy.
Estimating Error Rate: Collaborative vs. non-Collaborative

collaborative vs. non–collaborative estimation for $k=10$

- Mean absolute error of the MLE

- Number of unlabeled examples

- Guy Lebanon

Unsupervised Supervised Learning
Estimating Error Rate: USL vs. Supervised Learning

Guy Lebanon

Unsupervised Supervised Learning
Multiple annotators in two Amazon Mechanical Turk NLP tasks: textual entailment, and temporal event recognition
Estimating Error Rate: Ringnorm Dataset

Guy Lebanon | Unsupervised Supervised Learning
Estimating Error Rate: Amazon Product Reviews and 20-Newsgroups

<table>
<thead>
<tr>
<th>Domain Adapation (different product domains) and 20-newsgroups</th>
<th>book</th>
<th>dvd</th>
<th>kitchen</th>
<th>electronics</th>
<th>20newsgroup</th>
</tr>
</thead>
<tbody>
<tr>
<td>training error</td>
<td>0.22</td>
<td>0.23</td>
<td>0.26</td>
<td>0.30</td>
<td>0.028</td>
</tr>
<tr>
<td>non-collaborative</td>
<td><strong>0.04</strong></td>
<td><strong>0.04</strong></td>
<td><strong>0.08</strong></td>
<td><strong>0.06</strong></td>
<td><strong>0.006</strong></td>
</tr>
<tr>
<td>collaborative</td>
<td>0.10</td>
<td>0.10</td>
<td>0.09</td>
<td>0.08</td>
<td>n/a</td>
</tr>
</tbody>
</table>

Guy Lebanon
Unsupervised Supervised Learning
What’s Next?

- Estimate MSE for regression models (assuming a known $p(y)$)
- Fit regression models
- Structured Prediction
- Non-linear Kernel SVM
- Extension for Semi-supervised learning
- Extension for Active learning
What’s Next?

- Estimate MSE for regression models (assuming a known $p(y)$)
- Fit regression models
  - Structured Prediction
  - Non-linear Kernel SVM
  - Extension for Semi-supervised learning
  - Extension for Active learning
Estimate MSE for regression models (assuming a known $p(y)$)

- Fit regression models
- Structured Prediction
  - Non-linear Kernel SVM
  - Extension for Semi-supervised learning
  - Extension for Active learning
What’s Next?

- Estimate MSE for regression models (assuming a known $p(y)$)
- Fit regression models
- Structured Prediction
- Non-linear Kernel SVM
  - Extension for Semi-supervised learning
  - Extension for Active learning
What’s Next?

- Estimate MSE for regression models (assuming a known \( p(y) \))
- Fit regression models
- Structured Prediction
- Non-linear Kernel SVM
- Extension for Semi-supervised learning
- Extension for Active learning
What’s Next?

- Estimate MSE for regression models (assuming a known $p(y)$)
- Fit regression models
- Structured Prediction
- Non-linear Kernel SVM
- Extension for Semi-supervised learning
- Extension for Active learning
Thanks!

Collaborators:
- Krishnakumar Balasubramanian (Georgia Tech)
- Pinar Donmez (CMU → Yahoo)