CSC2412: Exponential Mechanism & Private PAC Learning

Sasho Nikolov
Classification Basics
The learning problem

**Problem**: develop an algorithm that classifies avocados into ripe and unripe.

We have a big data set of avocado data. For each avocado, we have:

- colour, firmness, size, shape, skin texture, ...
- ripe or not

From this data, we want to classify unseen avocados.

Learn a rule to predict label from features.
The learning problem, formally

Model:

- Known data universe \( \mathcal{X} \) and an unknown probability distribution \( D \) on \( \mathcal{X} \)
- Known concept class \( C \) and an unknown concept \( c \in C \)

Goal: Learn \( c \) from \( X \).

Assumption (realizability): Some \( c \in C \) can produce all correct labels.

All allowed rules to map features to label. E.g. all functions that depend on \( \leq 3 \) features.

E.g. an avocado is ripe if colour = green and firmness = medium. An approximation of \( c \).
The goal, formally

The error of a concept $c' \in C$ is

$$L_{D,c}(c') = \mathbb{P}_{x \sim D}(c'(x) \neq c(x)).$$

We want an algorithm $M$ that outputs some $c' \in C$ and satisfies

$$\mathbb{P}(L_{D,c}(M(X)) \leq \alpha) \geq 1 - \beta.$$
Empirical risk minimization

**Issue:** We want to find \( \arg \min_{c' \in C} L_{D,c}(c') \), but we do not know \( D, c \).

\( \overline{\text{In approximate minimizers are also ok}} \)

**Solution:** Instead we solve \( \arg \min_{c' \in C} L_X(c') \), where

\[
L_X(c) = 0
\]

is the empirical error.

\[
L_X(c') = \frac{|\{i : c'(x_i) \neq c(x_i)\}|}{n}
\]

fraction of pts in \( X \) misclassified by \( c' \)

Theorem (Uniform convergence)

Suppose that \( n \geq \frac{\ln(|C|/\beta)}{2\alpha^2} \). Then, with probability \( \geq 1 - \beta \),

\[
\max_{c' \in C} L_{D,c}(c') - L_X(c') \leq \alpha.
\]

Hoeffding's inequality (exercise)

\( \text{Pop and emp. loss are close for } \forall c' \in C \)

\( L_{D,c} \) population loss (unknown)

\( L_X \) empirical loss (known)

Other versions for infinite \( C \), e.g. VC-dimension
In private PAC learning, we require that

- when $X$ is a sample of iid labeled data points, we learn the correct concept, as in standard PAC learning;
- the learning algorithm is $\varepsilon$-differentially private for any labeled data set $X \in (\mathcal{X} \times C)^n$.

$$\forall X, X', \text{neighbouring} \quad \forall S \subseteq C$$

$$P(\mu(X) \in S) \leq e^\varepsilon \cdot P(\mu(X') \in S)$$

Privacy must hold even if data $X$ not iid
Want to do ERM w/ $\varepsilon$-DP i.e. (approximately) minimize

$$L_x(c') = \frac{\left| \{i : c(x_i) \neq c'(x_i) \} \right|}{n} \text{ over } c' \in C$$

How can we use Laplace noise mechanism for this?

$L_x(c')$ is a counting query

We could release answers to all counting queries $c = \frac{1}{k} e_1, \ldots, e_k$,

$$\frac{1}{k} L_x(c_1), L_x(c_2), \ldots, L_x(c_k)$$

Exercise: analyze this
Exponential mechanism
Private ERM

We want to solve \( \arg \min_{c' \in C} L_X(c') \).

How do we minimize with differential privacy?

Sample concepts with less error with higher probability

\[ \mathbb{P}(M(X) = c') \propto \exp \left( -\frac{\varepsilon n}{2} L_X(c') \right) \]

proportional to
Exponential Mechanism

**General set-up:** score function $u : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$

**Goal:** given $X$, find $\arg \max_y u(X, y)$

**Sensitivity**

$$\Delta u = \max_{y \in \mathcal{Y}} \max_{X \sim X'} |u(X, y) - u(X', y)|.$$  

The mechanism $M_{\text{exp}}(X)$ which outputs a random $Y$ so that

$$P(Y = y) = \frac{e^{\varepsilon u(X, y)/2\Delta u}}{\sum_{z \in \mathcal{Y}} e^{\varepsilon u(X, z)/2\Delta u}}$$

is $\varepsilon$-differentially private.
Privacy analysis

\[ P(Y = y) = \frac{e^{\varepsilon u(X, y) / 2\Delta u}}{\sum_{z \in \mathcal{Y}} e^{\varepsilon u(X, z) / 2\Delta u}} \]

\[
\begin{align*}
    \text{Enough to show: } & \forall x \sim x' \quad \forall y \in \mathcal{Y} \quad \frac{P(M(x) = y)}{P(M(x') = y)} \leq e^\varepsilon \\
    \frac{P(M(x) = y)}{P(M(x') = y)} & = \exp\left(\frac{\varepsilon (u(x, y) - u(x', y))}{2\Delta u}\right) \\
    & \leq e^{\varepsilon / 2} \cdot e^{\varepsilon / 2} = e^\varepsilon
\end{align*}
\]
Accuracy of the exponential mechanism

\[
\text{OPT}(X) = \max_{y \in Y} u(X, y)
\]

Then, for the output \( Y = M_{\text{exp}}(X) \),
\[
P(u(X, Y) \leq \text{OPT}(X) - t) \leq e^{\frac{\epsilon(\text{OPT}(X) - t)}{2\Delta u}} \cdot \sum_{z \in Y} e^{\frac{\epsilon u(X, z)}{2\Delta u}} \geq e^{\frac{\epsilon u(X, y^*)}{2\Delta u}}
\]

\[
\leq e^{\frac{\epsilon(\text{OPT} - t)}{2\Delta u}} \cdot \left(1 - \frac{1}{e^\epsilon}\right) \leq e^{-\frac{\epsilon t}{2\Delta u}} \cdot \left(1 - \frac{1}{e^\epsilon}\right)
\]

\[\text{y^* achieves OPT}(X)\]
Private Learning
Unknown distribution \( D \) on known \( \mathcal{C} \)

Unknown \( c \) in a known concept class \( \mathcal{C} \)

Data set \( X = \{(x_1, c(x_1)), \ldots, (x_n, c(x_n))\} \) where \( x_1, \ldots, x_n \sim_{\text{iid}} D \)

\[
L_{D, c} (c') = \mathbb{P}_{x \sim D} \left( c(x) \neq c'(x) \right) \\
L_{X} (c') = \frac{|\{i : c(x_i) \neq c'(x_i)\}|}{n}
\]

If \( n \geq \frac{\ln(1/C/\beta)}{2\Delta^2} \) then w/ prob \( \geq 1 - \beta \),
\[
\forall c' \in \mathcal{C} : L_{D, c} (c') \leq L_{X} (c') + \Delta
\]

**Exponential mechanism:** sample \( y \in \mathcal{Y} \) w/ prob. proportional to \( \exp(\epsilon u(x, y)/\Delta u) \)

\( \Delta u = \max_{y \in \mathcal{Y}} \max_{x \sim x'} |u(x, y) - u(x', y)| \)
Putting things together

A concept class $C$ can be learned by an $\epsilon$-differentially private mechanism when the sample size is

$$n \geq \max \left\{ \frac{4 \ln(2|C|/\beta)}{\epsilon \alpha}, \frac{2 \ln(2|C|/\beta)}{\alpha^2} \right\}$$

Use exp. mechanism with $G = C$

with $u(X, c') = -L_X(c')$

With prob $\geq 1 - \frac{\beta_2}{2}$, c' output by $\text{Map}(X)$ has $L_X(c') \leq \frac{\alpha}{2}$. By unit. convergence, w/ prob $\geq 1 - \frac{\beta_2}{2}$, $L_{D, c}(c') - L_X(c') \leq \frac{d}{2}$.
Putting things together

\[ u(X, c') = -L_X(c') \; ; \; \Delta u = \frac{1}{n} \; ; \; \text{sample } c' \in C \text{ w/ prob prop to } \exp(-enL_X(c')/2) \]

\[ \text{OPT}(X) = \max_{c' \in C} -L_X(c') \]
\[ = -\min_{c' \in C} L_X(c') = 0 \]

\[ P(L_X(M(X)) \geq \frac{d}{2}) = P(u(X, M(X)) \leq \text{OPT}(X) - \frac{d}{2}) \leq e^{-\epsilon^2 n/4} \cdot |C| \leq \frac{\beta}{2} \; \text{ if } n \geq \frac{4\ln(21|C|/\beta)}{\epsilon^2} \]

+ by unif. conv. (with \( \frac{d}{2}, \frac{\beta}{2} \)) w/ prob \( 1 - \frac{\beta}{2} \), \( L_D(c) (M(X)) \leq L_X(M(X)) + \frac{d}{2} \leq 2\epsilon^2 \cdot \frac{d}{2} + \epsilon^2 \epsilon = \epsilon \).