CSC2412: Exponential Mechanism & Private PAC Learning

Sasho Nikolov

Classification Basics

The learning problem

+1 -1

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:

features

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

From this data, we want to classify unseen avocados.

Learn a rule to predict label from features

The learning problem, formally

Model: all possible setting to the features

- Known data universe $\boxed{\mathcal{X}}$ and an unknown probability distribution D on $\mathcal{X}$

- Known concept class C and an unknown concept $c \in C$

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 ≤ 3 features

- We get a dataset $X = \{(\underbrace{x_1}_{\text{features}}, \underbrace{c(x_1)}_{\text{label}}), \ldots, (x_n, c(x_n))\}$, where each x_i is an independent sample from D.

E.g. an avocado is ripe iff colour = ___ and firmness = medium } c

Goal: Learn an approximation of c from X.

The goal, formally

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

Fraction of the population labeled incorrectly by c'

$$\underbrace{L_{D,c}(c')}_{\text{loss of } c' \text{ (w.r.t. } D,\, c)} = \mathbb{P}_{x \sim D}(c'(x) \neq c(x)).$$

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

on input $X = ((x_1, c(x_1)), \ldots, (x_n, c(x_n)))$ sampled iid from D

$$\mathbb{P}(\underbrace{L_{D,c}(\mathcal{M}(X)) \leq \alpha}_{\text{output of } \mathcal{M}(X) \text{ misclassifies} \leq \alpha \text{ fraction of the population}}) \geq 1 - \beta.$$

$\mathbb{P}$: taken over randomness in choosing X and any randomness of $\mathcal{M}$

Probably Approximately Correct learning (PAC) [Valiant]

Empirical risk minimization

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

↳ approximate minimizers are also ok

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

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

$L_X(c) = 0$

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

fraction of pts in X misclassified by c'

is the empirical error.

L_X empirical loss (known)

Theorem (Uniform convergence) → Pop and emp. loss are close for $\forall c' \in C$

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

Hoeffding's inequality (exercise)

$$\max_{c' \in C} L_{D,c}(c') - L_X(c') \leq \alpha. \quad \forall c' \in C$$

$L_{D,c}(c') \leq L_X(c') + \alpha$

Other versions for infinite C, e.g. VC-dimension

Private learning

In private PAC learning, we require that

- when X is a sample of iid labeled data points, we learn ~~the~~ an approximately correct concept, as in stadard PAC learning;

  $\mathcal{M}$ that on input X outputs $c' \in C$

- the learning algorithm is ε-differentially private for **any** labeled data set $X \in (\mathcal{X} \times C)^n$.

Privacy must hold even if data X not iid

$\forall$ X, X' neighbouring $\forall S \subseteq C$

$$\mathbb{P}(\mathcal{M}(X) \in S) \leq e^{\varepsilon} \cdot \mathbb{P}(\mathcal{M}(X') \in S)$$

Want to do ERM w/ ε-DP

i.e. (approximately) minimize

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

over $c' \in \mathcal{C}$

How can we use Laplace noise mechanism for this?

$L_X(c')$ is a counting query

Exercise: analyze this

We could release answers to all counting queries

$$\{L_X(c_1), L_X(c_2), \ldots, L_X(c_k)\}$$

$\mathcal{C} = \{c_1, \ldots, c_k\}$

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

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

proportional to

Exponential Mechanism

output space

General set-up: score function $u : \mathcal{X}^n \times \mathcal{Y} \to \mathbb{R}$

$u(X, y)$ = "how good of an output is y for the dataset X?"

Goal: given X, find $\arg\max_{y \in \mathcal{Y}} u(X, y)$

Sensitivity

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

How different can the score be between neighbouring X, X'

The mechanism $\mathcal{M}_{\exp}(X)$ which outputs a random Y so that

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

$\to$ normalizing factor

is ε-differentially private

Privacy analysis

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

Enough to show: $\forall X \sim X'$, $\forall y \in \mathcal{Y}$ $\quad \dfrac{\mathbb{P}(\mathcal{M}(X) = y)}{\mathbb{P}(\mathcal{M}(X') = y')} \leq e^{\varepsilon}$

$$\frac{\mathbb{P}(\mathcal{M}(X) = y)}{\mathbb{P}(\mathcal{M}(X') = y)} = \exp\left(\frac{\varepsilon(u(X,y) - u(X',y))}{2\Delta u}\right) \cdot \frac{\sum_{z \in \mathcal{Y}} e^{\varepsilon u(X',z)/2\Delta u}}{\sum_{z \in \mathcal{Y}} e^{\varepsilon u(X,z)/2\Delta u}}$$

$$\leq e^{\varepsilon/2} \cdot e^{\varepsilon/2} = e^{\varepsilon}$$

Accuracy of the exponential mechanism

≥ 1 output y^* achieves $\underline{\mathrm{OPT}(X)}$

$$\mathrm{OPT}(X) = \max_{y \in \mathcal{Y}} u(X, y)$$

Then, for the output $Y = \mathcal{M}_{\exp}(X)$,

$$\mathbb{P}(u(X, Y) \leq \mathrm{OPT}(X) - t) \leq \frac{e^{\varepsilon(\mathrm{OPT}(X) - t)/2\Delta u}}{\sum_{z \in \mathcal{Y}} e^{\varepsilon u(X,z)/2\Delta u}} \cdot |\{y : u(X,y) \leq \mathrm{OPT} - t\}|$$

$$\geq e^{\varepsilon \underline{u(X,y^*)}/2\Delta u}$$

$$\leq \frac{e^{\varepsilon(\mathrm{OPT} - t)/2\Delta u} \cdot (|\mathcal{Y}| - 1)}{e^{\varepsilon \mathrm{OPT}/2\Delta u}} \leq e^{-\varepsilon t/2\Delta u} \cdot |\mathcal{Y}|$$

$$\leq e^{-\varepsilon t/2\Delta u} \cdot |\mathcal{Y}|$$

Private Learning

Unknown distribution D on known $\mathcal{X}$

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

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

$$L_{D,c}(c') = \mathbb{P}_{x \sim D}(c(x) \neq c'(x)) \quad \Big| \quad L_X(c') = \frac{|\{i : c(x_i) \neq c'(x_i)\}|}{n}$$

If $n \geq \dfrac{\ln(|\mathcal{C}|/\beta)}{2\alpha^2}$ then w/ prob $\geq 1-\beta$,

$$\forall c' \in \mathcal{C} : L_{D,c}(c') \leq L_X(c') + \alpha$$

Exponential mechanism: sample $y \in \mathcal{Y}$ w/ prob.

proportional to $\exp(\varepsilon u(X,y)/2\Delta u)$

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

Putting things together

A concept class C can be learned by an ~~eps~~ ε-differentially private mechanism when the sample size is

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

Use exp. mechanism with $\mathcal{Y} = C$

with $\boxed{u(X, c') = -L_X(c')}$

With prob $\geq 1 - \beta/2$, c' output by $M_{exp}(X)$ has $L_X(c') \leq \frac{\alpha}{2}$. By unif. convergence, w/ prob $\geq 1 - \beta/2$, $L_{D,c}(c') - L_X(c') \leq \frac{\alpha}{2}$

Putting things together

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

$\downarrow$

ε-DP

$$OPT(X) = \max_{c' \in C} -L_X(c')$$

$$= -\min_{c' \in C} L_X(c') = 0$$

$$\mathbb{P}\left(L_X(\mathcal{M}(X)) \geq \frac{\alpha}{2}\right) = \mathbb{P}\left(u(X, \mathcal{M}(X)) \leq OPT(X) - \frac{\alpha}{2}\right)$$

$$\leq e^{-\varepsilon \alpha n/4} \cdot |C| \leq \beta/2 \quad \text{if} \quad n \geq \frac{4\ln(2|C|/\beta)}{\varepsilon \alpha}$$

+ by unif. conv. (with $\alpha/2$, $\beta/2$) w/ prob $1 - \beta/2$, $L_{D,c}(\mathcal{M}(X)) \leq L_X(\mathcal{M}(X)) + \alpha/2 \underset{\text{w/prob} \geq 1-\beta}{\leq} \alpha/2 + \alpha/2 = \alpha$.