More Generalization Theorems

Yufei Tao

Department of Computer Science and Engineering
Chinese University of Hong Kong

Classification (Re-defined on Numeric Attributes)

Define the **universe** to be $\mathbb{R}^d$, where $\mathbb{R}$ is the set of real values.

Each **object** is a point p of $\mathbb{R}^d \times \{-1, 1\}$, i.e., it has coordinate $p[i]$ on the i-th $(1 \leq i \leq d)$ dimension of the universe, and a **class label** $p[C]$ either -1 or 1.

Denote by D a probabilistic distribution on $\mathbb{R}^d \times \{-1, 1\}$.

Classification (Re-defined on Numeric Attributes)

Goal: Given an object p drawn from D, we want to predict its label $p[C]$ from its coordinates $p[1], ..., p[d]$.

We do so by constructing a function

$$M : \mathbb{R}^d \rightarrow \{-1, 1\}$$

which we refer to as a **classifier**. Given any point p, we predict its class label as $M(p[1], ..., p[d])$.

We define the error of M on D—denoted as $err_D(M)$—as:

$$err_D(M) \quad = \quad \boldsymbol{Pr}_{p \sim D}[M(p[1], ..., p[d]) \neq p[C]].$$

Ideally, we want to find an M to minimize $err_D(M)$.

Classification (Re-defined on Numeric Attributes)

In training, we are given a sample set P_{train} of D, where each object in P_{train} is drawn independently according to D. We refer to P_{train} as the **training set**.

We would like to learn our classifier M from P_{train}.

Henceforth, let us focus on an (arbitrary) set $\mathcal{M}$ of classifiers.

Example: Let us define a **linear classifier** as a function M that is defined by co-efficients $c_1, ..., c_d, c_{d+1}$ such that

- $M(p) = 1$ if $\sum_{i=1}^{d} c_i \cdot p[i] \geq c_{d+1}$;
- $M(p) = -1$ if $\sum_{i=1}^{d} c_i \cdot p[i] < c_{d+1}$.

In this lecture we will be particularly interested in the set $\mathcal{M}$ of all linear classifiers.

Shattering

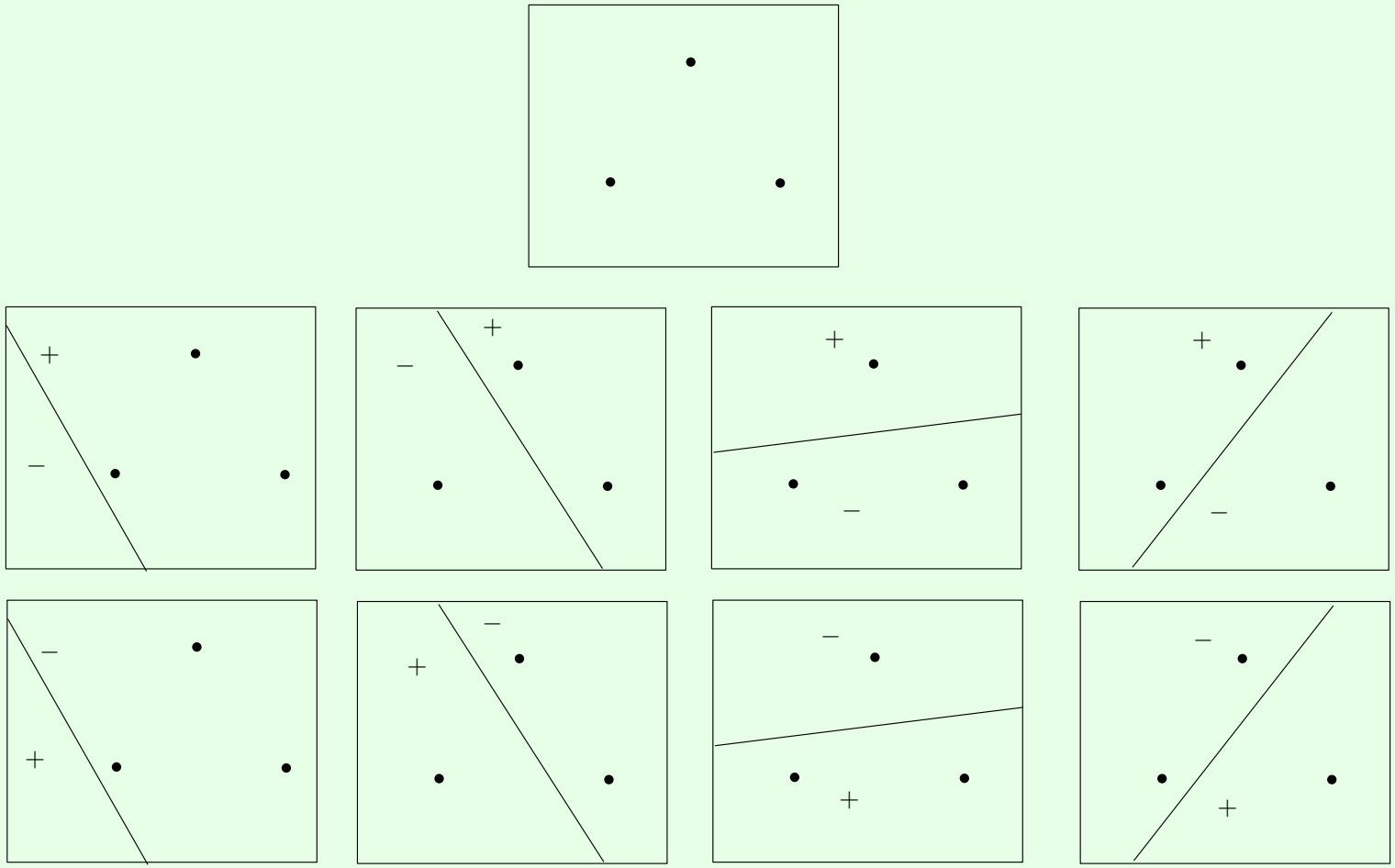
Let P be a set of points in $\mathbb{R}^d$. Given a classifier $M \in \mathcal{M}$, we define:

$$P_M = \{p \in P \mid M(p) = 1\}$$

that is, the set of points in P which are mapped to label 1 by M.

We say that $\mathcal{M}$ **shatters** P if, for any subset $S \subseteq P$, there exists a classifier $M \in \mathcal{M}$ such that $S = P_M$.

Example: Suppose that $\mathcal{M}$ is the set of linear classifiers for $d = 2$. It can shatter the following the set P of three points as shown below.



Example (cont.): Can you find 4 points in $\mathbb{R}^2$ that can be shattered by $\mathcal{M}$?

The answer is **NO**! Can you prove this?

VC Dimension

The **VC-dimension** of a set $\mathcal{M}$ of classifiers on a set P of points is the size of the largest point set P that can be shattered by $\mathcal{M}$.

Note: The VC-dimension is defined on the pair $(P, \mathcal{M})$. In other words, even for the same $\mathcal{M}$, the VC-dimension may change when P varies.

VC Dimension of the Set of Linear Classifiers

Theorem: The VC-dimension of the set $\mathcal{M}$ of linear classifiers on $\mathbb{R}^d$ (i.e., the set of all d-dimensional points) is $d + 1$.

The proof is beyond the scope of the course.

Example: We have seen earlier that when $d = 2$, the set $\mathcal{M}$ of linear classifiers can shatter **at least one** set of 3 points, but cannot shatter **any** set of 4 points; hence, the VC-dimension of $(\mathbb{R}^2, \mathcal{M})$ is 3.

VC-Based Generalization Theorem

The **support set** of the underlying distribution D is the set of points in $\mathbb{R}^d$ that have a positive probability to be drawn according to D.

Theorem: Let P be the support set of D, and $\mathcal{M}$ be a (possibly infinite) set of classifiers each of which maps a point in P to $\{-1, 1\}$. Suppose that the VC-dimension of $(P, \mathcal{M})$ is λ. Given any value $0 < \delta \leq 1$, it holds with probability at least $1 - \delta$ that, for **every** $M \in \mathcal{M}$,

$$err_D(M) \leq err_R(M) + \sqrt{\frac{8 \ln \frac{4}{\delta} + 8\lambda \cdot \ln \frac{2e|P_{train}|}{\lambda}}{|P_{train}|}}.$$

where P_{train} is the set of training points.

The proof is beyond the scope of the course.

To see the usefulness of the new generalization theorem, notice that it places **no constraints** on how many bits are used to encode a classifier. In fact, there are **infinitely many** linear classifiers, which means that they cannot be represented using a finite number of bits! Therefore, the new generalization theorem still works even when the old one fails.

In particular, think about what implications you can drawn about the Perceptron algorithm (we no longer need to discuss how many bits are needed to represent the coefficients).

The new generalization theorem also suggests that if a set $\mathcal{M}$ of classifiers is "**powerful**" — namely it can shatter a large number of points — it is **more difficult** to learn in the sense that a larger amount of training data is needed.

Even for the set $\mathcal{M}$ of linear classifiers, the size of the training set P_{train} needs to be $\Omega(d)$ to ensure a small generalization error (remember that the VC-dimension of $\mathcal{M}$ is $d+1$). This would become a problem when d is large: in fact, in some situations we even have to push d to ∞, as will seen later in the course.

Next, we will introduce yet another generalization theorem that works only on **linearly separable** training data.

Let us re-define a **linear classifier** as a function M that is defined by a d-dimensional vector $\vec{c} = (c_1, ..., c_d)$ such that, given a point $\vec{p} = (p[1], ..., p[d])$,

- $M(p) = 1$ if $\vec{c} \cdot \vec{p} \geq 0$;
- $M(p) = -1$ if $\vec{c} \cdot \vec{p} < 0$.

Let P_{train} be the set of training points. Suppose that P_{train} is linearly separable. This means that there exists some linear classifier defined by $\vec{c}$ such that, for every point $p \in P_{train}$:

- $\vec{c} \cdot \vec{p} > 0$ if p has label 1
- $\vec{c} \cdot \vec{p} < 0$ if p has label -1,

Define $R = \max_{p \in P_{train}} |\vec{p}|$
namely R is the maximum distance between a point in the training set and the origin.

Let M be any linear classifier that **separates** the points in P_{train} of the two labels. This means that for every point $p \in P_{train}$:

- $\vec{c} \cdot \vec{p} > 0$ if p has label 1
- $\vec{c} \cdot \vec{p} < 0$ if p has label -1

where $\vec{c}$ is the coefficient vector of M.

We say that M is **canonical** if for every point $p \in P_{train}$:

- $\vec{c} \cdot \vec{p} \geq 1$ if p has label 1
- $\vec{c} \cdot \vec{p} \leq -1$ if p has label -1;

and the equality holds on **at least one point** in P_{train}.

Every linear classifier can be written into a **unique** canonical form.

Think: why?

Margin-Based Generalization Theorem

Theorem: Suppose that the set P_{train} of training points is **linearly separable**. Given any value $0 < \delta \leq 1$, it holds with probability at least $1 - \delta$ that, for **every canonical** $M \in \mathcal{M}$ separating P_{train},

$$err_D(M) \leq \frac{4R \cdot |\vec{c}|}{\sqrt{|P_{train}|}} + \sqrt{\frac{\ln \frac{2}{\delta} + \ln \lceil \log_2 |\vec{c}| \rceil}{|P_{train}|}}.$$

where $\vec{c}$ is the coefficient vector of M.

The proof is beyond the scope of the course.

Note that the theorem does not depend on the dimensionality d of the data space.

Margin-Based Generalization Theorem

Why is the theorem "margin-based"?
The margin of the classifier M equals $1/|\vec{c}|$ — try to derive this yourself (we will do so explicitly later in the course).

The theorem has an important implication: when the training set P_{train} is linearly separable, we should not be content with just finding any linear classifier to separate P_{train}, but should strive to find the one with the **largest** margin. This is precisely what we will do in the next lecture.