Passive vs. Active Learning

Huan-Kai Tseng

College of Social Sciences
National Taiwan University

January 3rd, 2019

Table of contents

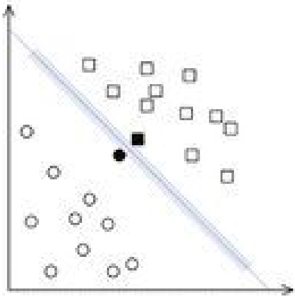
SVM basics

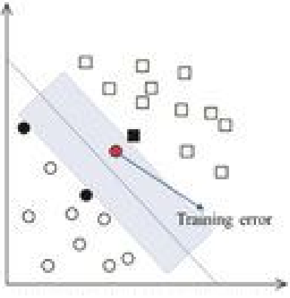
Active Learning

Further readings

Recap: SVM

Figure 1: Separating two classes



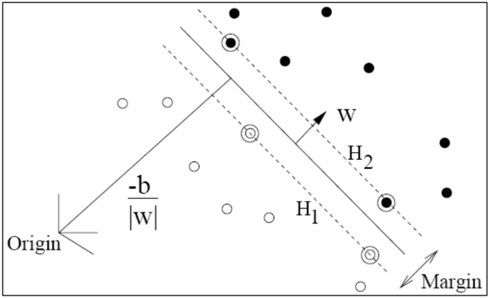


For a set of data points in $\mathbb{R}^d$, we want to find a **support hyperplane** (SHP) that perfectly classifies two classes: □, ○.

Contrary to s.e., we want the SHP to be as wide as possible such that we are more confident in separating these two classes (i.e., producing less uncertain classes, such as ●).

Recap: SVM

Figure 2: Maximizing SHP



Specifically, the training task comes down to finding a SHP that maximizes the distance between H1 and H2, a space that perfectly classifies any data points from the training data $\{x_i, y_i\}$ for $i = 1, 2, \ldots, n$, $x_i \in \mathbb{R}^d$, and $y_i \in \{b, -b\}$.

Recap: SVM

Let **Classification Hyperplane** be $w^T = -b(w^T + b = 0)$, SHP can be re-written as

$$H_1 : w^T x + b + \delta$$
$$H_1 : w^T x + b - \delta$$

Fixing δ to a constant

$$H_1 : w^T x + b = 1$$
$$H_1 : w^T x + b = -1$$

The distance from H1 (H2) to the origin:

$$\frac{|1-b|}{\|w\|}$$
$$\frac{|-1-b|}{\|w\|}$$

Recap: SVM

For x_i in $\mathbb{R}^d$, they must satisfy

$$H_1 : w_i^T x_i + b \geqslant 1, \text{if } y_i = 1$$
$$H_1 : w_i^T x_i + b \leqslant 1, \text{if } y_i = -1$$

which is reduced to $y_i \left(w_i^T x_i + b \right) \geqslant 1$.
Our objective is to maximize $\frac{2}{\|w\|}$ or minimize $\frac{\|w\|}{2}$.
The above questions can be summarized as

$$\begin{cases} min_{w,b} \frac{1}{2} w^T w \\ y_i \left(w_i^T x_i + b \right) \geqslant 1. \end{cases} \tag{1}$$

Using Lagrangian to solve for (1)

$$\mathscr{L}(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i}^{n} \alpha_i \left[y_i \left(w_i^T x_i + b \right) - 1 \right], \tag{2}$$

satisfying min_w, max_α, $\mathscr{L}(w, \alpha)$.

Recap: SVM

Differentiating (2) wrt.

$$w : w^* \sum_i \alpha_i y_i x_i$$

$$b : \sum_i \alpha_i y_i$$

substituting the above conditions into (2), we get the duality

$$\text{Max} \mathscr{L} = \sum_i \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$\text{s.t.} \sum_i \alpha_i y_i$$

$$\alpha_i \geqslant 0 \text{ for all } i$$

Since all x_i, x_j in $\mathbb{R}^d$ (where $x_i \neq x_j$) are expressed in inner product form, $x_i^T x_j$, the SVM can be solved using quadratic programming (e.g., kernel).

Recap: SVM

Since

$$y_i = \underbrace{\sum_i \alpha_i y_i x_i^T x_j}_{w^T x} + b$$

rearranging

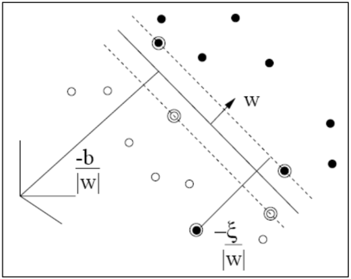
$$-b = \sum_i \alpha_i y_i x_i^T x_j - y_i$$

We know that

- each (x_i, y_i) in $\mathbb{R}^d$ defines a “support vector” on the SHP
- any two support vectors determine w and b
- by (2), only when $\alpha_i > 0$ do (x_i, y_i) affect w and b

Dealing with misclassification

Figure 3: Classification error



How do we deal with misclassified cases?

1. Adding an error term ξ
2. Consult human experts: **active learning**

Dealing with misclassification

The sources of error

- **Rare events**: minority groups, extremely rare occurrences
  - restaurant ratings of Ethiopian restaurants within Da-an district given by NTU sophomores majoring PolSci
- Your algorithms have a hard time ascertaining the **class(es)** of particular cases
  - I ain't know nothing ⇒ you know something or nothing?
  - 阿不就好棒棒 ⇒ 到底是棒還是不棒?

Dealing with misclassification

Consult human experts

Dealing with misclassification

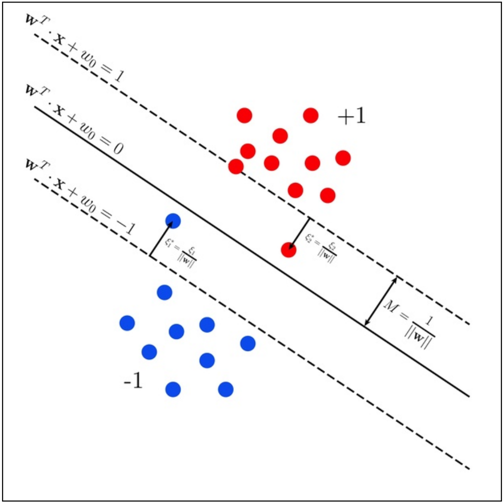
Why do we need human experts?
What for?
How can the machine learn from human experts?

Dealing with misclassification

Passive Learning: Experts begins by labeling a sample of documents according to some predefined concepts. This subset of documents is then used to train a learning algorithm to automatically predict the label of new documents outside of training set. This approach breaks down when labeling documents is costly or the concepts being measured are extremely rare.

Active Learning: Analysts would iteratively answer a learning algorithm's queries for documents that the model is most uncertain about. After labeling a single document or batch of documents, a new learning algorithm will be trained and will execute a new query of unlabeled documents for the expert coder to label [so on and so forth...]

Dealing with misclassification


$\mathbf{w}^T \cdot \mathbf{x} + w_0 = 1$
$\mathbf{w}^T \cdot \mathbf{x} + w_0 = 0$
$\mathbf{w}^T \cdot \mathbf{x} + w_0 = -1$
+1
$\xi_1^* = \frac{\xi_1}{||\mathbf{w}||}$
$\xi_2^* = \frac{\xi_2}{||\mathbf{w}||}$
$M = \frac{1}{||\mathbf{w}||}$
-1

Pick the data points the algorithm is most uncertain about

Dealing with misclassification

Procedure

1. Start with an initial set of documents $x_i \in \boldsymbol{X}$ with known labels $\mathbf{y}$
2. Train the algorithm using training data sampled from $(\boldsymbol{X}, \mathbf{y})$
3. Produce a predicted probability for each unknown document: $\hat{\mathbf{y}}^* = f(\mathbf{X}^*, \theta^*)$
4. Use the query function to obtain a new document for labeling: $z = q(\mathbf{y}^*) = \text{argmin}_i |\hat{y}^* - 0.5|$
5. Obtain a label y_z for x_z from human experts
6. Replace $x_z \in \mathbf{X}^*$ with x_z that you just labelled
7. Repeat Steps 2-5 until a stopping criterion is reached

We should get a new learning algorithm with improved classification accuracy!

Dealing with misclassification

Implications

- An online crowdsourcing active learning platform
- Human labelers with different cultural backgrounds
- Inter-coder reliability

Further readings

Settles, Burr. 2012. Active Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning 6(1): 1-114.

Araújo, Priscilla. 2018. “Do You Know What a Data Labeler Does?” *Towards Data Science* Feb. 2nd, 2018.