Ex SVM for binary classification

SVM = Support Vector Machine

Goal: give training examples $\{(x_i, y_i)\}_{i=1}^{n}$

where $x_i \in \mathbb{R}^d$ feature vectors

$y_i \in \{-1, +1\}$ class labels

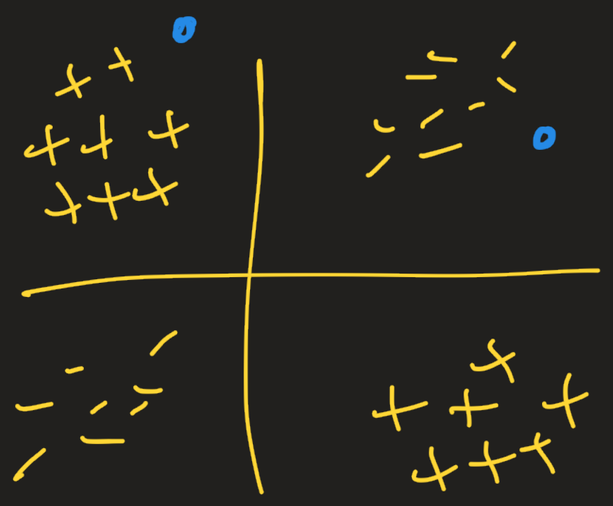
want to learn a classification function

$$f : \mathbb{R}^d \to \{-1, +1\}$$

Motivating Example: mortgage default determination

$$x = \begin{bmatrix} \text{credit score} \\ \text{debt-to-income ratio} \\ \text{percentage paid down} \\ \text{first loan?} \\ \text{mortgage insurance?} \\ \vdots \end{bmatrix}$$

Model: $y = \text{sign}(\langle w, x \rangle + b)$ $O(d)$ storage



k-nn

$O(nd)$ storage

Optimization Problem

- We want $\text{sign}(\langle w, x_i \rangle + b)$ to match y_i for all our training data

$$\Leftrightarrow \quad y_i(\langle w, x_i \rangle + b) > 0$$

- In fact, we want the x_i to be far away from the decision surface

$$\Leftrightarrow \quad y_i(\underbrace{\langle w, x_i \rangle + b}) > 1$$

so our predictions are confident/robust

- Be careful: we can trivially get confidence just by scaling up w

Consider the case $b=0$, then the confidence of the model for the training data is

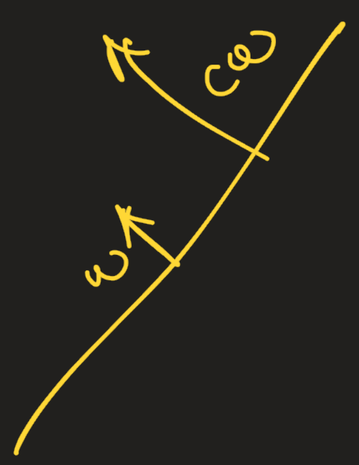
$$\min |\langle w, x_i \rangle| = \varepsilon$$

the smallest distance of any feature vector to the classification surface

I can rescale to $\tilde{w} = \frac{w}{\varepsilon}$, then

$$\min |\langle \tilde{w}, x_i \rangle| = 1$$

But the model is the same, geometrically



where c is some scaling factor > 0

- To avoid this trivial scaling, we require $\|w\|_2$ to be small

$\left(\|x\|_2 = \sqrt{\sum x_i^2} \right)$

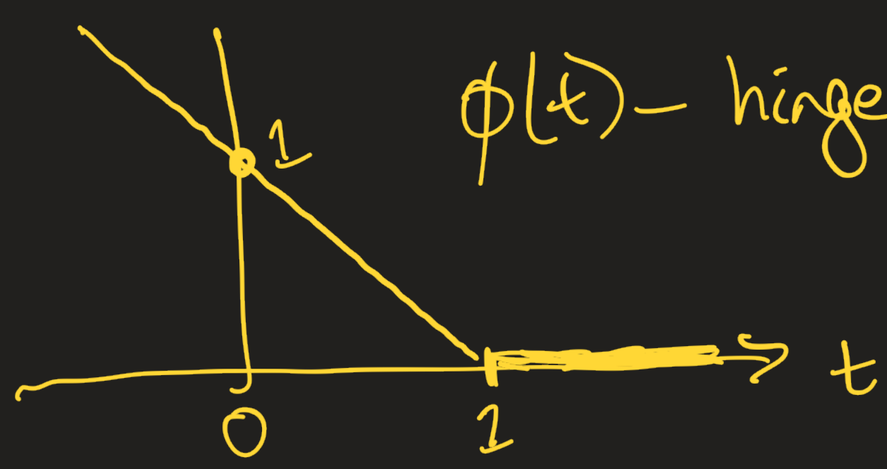
That leads to the optimization problem

$$\min_{w,b} \frac{1}{n} \sum \phi(y_i(\langle w, x_i \rangle + b)) + \frac{\lambda}{2}\|w\|_2^2$$

where ϕ is chosen to encourage HS input to be large and positive

for SVMs we take $\phi(t) = \max(1-t, 0)$

$= (1-t)_+$



$\phi(t)$ – hinge loss

Solutions to the opt prob

1) duality (skip)

2) gradient methods

$$\min_{w} f(w) \;\overset{\text{TS}}{\approx}\; \min_{w} \; f(w_k) + \nabla f(w_k)^T (w - w_k) + \underbrace{\text{h.o.t}}_{\text{higher order terms}}$$

(w_k: current approx of a minimizer)

take a small step in the negative gradient direction:

$$w_{k+1} = w_k - \mu \nabla f(w_k)$$

$$-\mu \nabla f(w_k) = w_{k+1} - w_k$$

$$\Rightarrow f(w_{k+1}) \approx f(w_k) + \nabla f(w_k)^T (w_{k+1} - w_k)$$

$$= f(w_k) - \mu \|\nabla f(w_k)\|_2^2$$

$$< f(w_k)$$

$$-\mu \nabla f(w_k)^T \nabla f(w_k) = -\mu \|\nabla f(w_k)\|$$

In our case we're minimizing (let $b=0$)

$$f(w) = \frac{1}{n}\sum_{i=1}^{n} \phi(y_i \langle w, x_i \rangle) + \frac{\lambda}{2}\|w\|_2^2$$

so

$$\nabla f(w) = \lambda w + \frac{1}{n}\sum_{i=1}^{n} \begin{cases} y_i x_i & \text{if } \phi(y_i \langle w, x_i \rangle) > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\nabla_w \phi(y_i \langle w, x_i \rangle) = \phi'(y_i \langle w, x_i \rangle)\, y_i x_i$$

$$\phi'(t) = \begin{cases} -1 & \text{if } \phi(t) > 0 \\ 0 & \text{if } \phi(t) = 0 \end{cases}$$

point: evaluating $\nabla f(\omega_k)$ takes $O(nd)$

and if we want to get close to the optimal model ω^*:

$$f(\omega_k) \leq f(\omega_*) + \epsilon$$

need $k = O\left(\frac{1}{\lambda}\log\left(\frac{1}{\epsilon}\right)\right)$ $\Rightarrow$ implies we have an error that looks like

$$f(\omega_T) \leq f(\omega_*) + O(e^{-T})$$

after T steps

For reasons we won't go into $\lambda \approx \frac{1}{\sqrt{n}}$ for optimal behavior

$\Rightarrow$ overall we need $O\left(nd \times \sqrt{n}\log\left(\frac{1}{\epsilon}\right)\right)$ operations to reach ϵ-suboptimality

When n is large, this $O(n^{3/2})$ dependence is prohibitive.

Solution: don't use the full gradient.

(stochastic first-order methods)

Lightning review of Prob theory

- What is prob theory?
  - a mathematical framework for dealing w/ uncertainty & randomness
  - ML = probability + optimization + modeling
- Why is it relevant to this course?

Basics of prob theory

$\Omega = \{\omega_1, \dots\}$ is a set called our universe, sample space

ω_1 ← elementary events

p is a probability distribution over Ω if

$p: \Omega \to [0,1]$ ← (true only if Ω is discrete; if Ω is not, p can be unbounded, but $p > 0$)

and $\sum_{\omega \in \Omega} p(\omega) = 1$

An event E is a subset of Ω. We define the prob of an event as

$$p(E) = \sum_{\omega \in E} p(\omega)$$

notice for all events $0 \leq p(E) \leq 1$

Ex $\Omega = \{1, 2, \cdots, 6\}$ corresp to roll of a die

$p(1) = \cdots = p(6)$ corresp to a fair die

$= \frac{1}{6}$

E = even numbers $= \{2, 4, 6\}$

$p(E) = p(2) + p(4) + p(6) = 3 \cdot \frac{1}{6} = \frac{1}{2}$

Ex $\Omega = \mathbb{R}^+ = \{x \mid x > 0\}$

$p(x) = \lambda e^{-\lambda x}$ for $\lambda > 0$

$\int_{\mathbb{R}^+} p(x)\,dx = -e^{-\lambda x}\Big]_0^\infty = -[0 - 1] = 1$

Ex $\Omega = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \right\}$

$p(\omega) = \frac{1}{4}$ for all ω

$E = \{ x \in \Omega : x \text{ has an even number of } 0\text{s} \}$

$= \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right\}$

$p(E) = \frac{1}{4} \cdot 2 = \frac{1}{2}$

Some calculus rules for probabilities of sets

$$p(\mathcal{Z}) = 1$$

$$p(\phi) = 0$$

$$p(E^c) = 1 - p(E)$$

- if $E_1, E_2, \ldots$ are mutually exclusive $\left(E_i \cap E_j = \phi \text{ if } i \neq j\right)$

$$p\left(\bigcup_i E_i\right) = \sum_i p(E_i)$$

- if $E_1 \subseteq E_2$

$$p(E_1) \leq p(E_2)$$

Random Variables

A random variable X takes its values in a set Ω_X according to a probability distribution $\mathbb{P}$ (or $\mathbb{P}_X$).

We say X is distributed according to $\mathbb{P}$, written as

$$X \sim \mathbb{P}$$

Think of $\{X = x\}$ as an elementary event that is either true or false and

$$\mathbb{P}(X = x) = \mathbb{P}_X(x)$$

is the probability of truth

Let $S = \{s_1, \dots\}$, then $\mathbb{P}(X \in S) = \mathbb{P}_X(S)$

$$= \sum_{s \in S} \mathbb{P}(X = s)$$

Ex X — a word in the English language

$P_X(w)$ = frequency of usage for w in Wikipedia

↑ some word

For random variables in an infinite set or a large finite set, we use <u>parametric models</u> for their distributions, that explain how $P_X(x)$ varies as a function of x and some parameters of the model.

Ex Bernoulli distribution on $\{0,1\}$: flipping a biased coin

write $X \sim \text{Bern}(p)$, p is the prob of 1

$P_X(1) = p$

$P_X(0) = 1-p$

Ex Binomial distribution, $Bin(n, p)$: flip biased coin n times, count the number of heads

$$\left.\begin{array}{l} P(0) = \\ \vdots \\ P(n) = \end{array}\right\} \Rightarrow P(i) = \binom{n}{i} p^i (1-p)^{n-i}$$

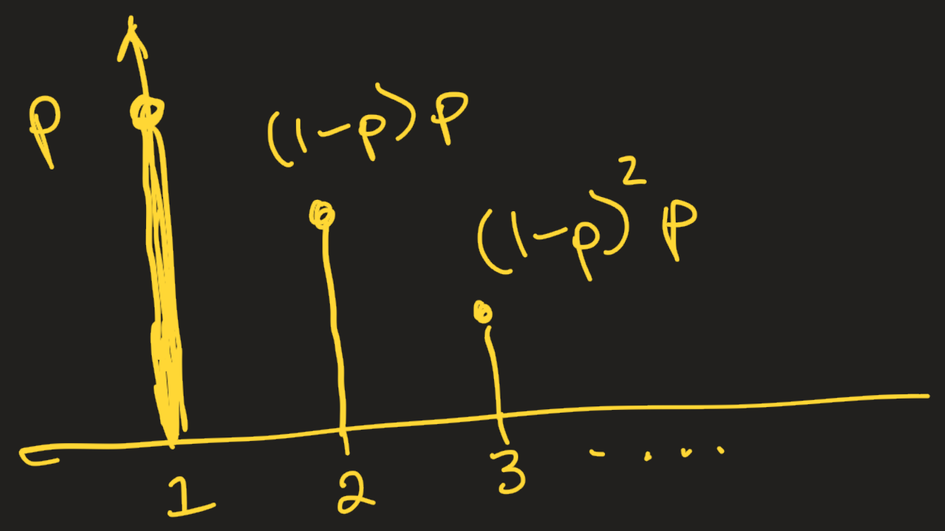
Discrete random variables (call $\mathbb{P}$, p the probability mass function)

the values $\mathcal{Z}_X$ can be enumerated (they are finitely many, or countably many)

Ex: Bernoulli, Binomial, Geometric "pmf"

Geometric r.v. $\text{Geom}(p)$: what is the number of flips at which the first head is observed?

$$\mathbb{P}(i) = (1-p)^{i-1} p \qquad i = 1, \dots$$



Continuous r.v.

A random variable that takes values in $\mathbb{R}$ and has a probability density function (pdf)

$$\mathbb{P}(a \leq X \leq b) = \int_a^b p(x)\, dx$$

where $p(x)$, the pdf, is nonnegative everywhere

Ex $U[a,b]$ has

$$p(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b \\ 0 & \text{elsewhere} \end{cases}$$

Ex Gaussian (or normal) r.v.,

$$X \sim \mathcal{N}(\mu, \sigma^2)$$ has density

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-1}{2\sigma^2}(x-\mu)^2\right]$$

Here: μ — "mean"

σ^2 — "variance"