LECTURE - 1

dt: 25/07/14

$\mathcal{D} = \{(x_1, y_1), (x_2, y_2) \ldots (x_m, y_m)\}$ $\quad x \in \mathbb{R}^N \quad y_i \in \{-1, 1\}$

<u>classificn</u>: Find a function $f(x)$ such that it explains y in the best possible manner

$\Rightarrow$ $f(x)$ should minimize the error/loss.

$$\underset{f}{\operatorname{argmin}} \, |\, l(y - f(x)) \,| \quad \text{where } l \text{ stands for "loss" function.} \longrightarrow ①$$

Now, if you set something like $f(x) = \begin{cases} y_i & \text{for } i = 1 \ldots m \\ 0 & \text{otherwise} \end{cases}$ that will minimize the error (=0) but it is nothing but memorizing everything and it will fail for unseen data.

Now, lets assume that x & y jointly come from a mixture ~~&~~ distribun $F_{xy}(x, y)$. Obviously, since random variable X is continuous & Y is discrete, marginalizn w.r.t x & y will give rise to a P.D.F and a P.M.F respectively.

$F_x = f_x$ (pdf)
$F_y = f_y$ (pmf)
$F_{X|Y} = f_{X|Y}$ (pmf)
$$F_{Y|X} = \lim_{\Delta x \to 0} \frac{P(x - \Delta x \le X < x + \Delta x,\ Y = y)}{P(x - \Delta x \le X < x + \Delta x)} \text{ (pdf)}$$
Notations

<u>Prove</u>: (Bayes theorem holds for Random variables as well.)

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y) \cdot f_Y(y)}{f_X(x)}.$$

Now an ideal expansion of equn (1) is to generalize by considering the expected loss function, i.e for an unknown distribution explaining the data, the function that minimizes the expected loss (Average weighted loss for an infinite number of samples) is the best function.

$$\underset{f(x)}{\operatorname{argmin}} = E[l(y - g(x))] = \iint_{\forall y\, \forall x} l(y, g(x)) \cdot dF_{xy} \cdot dy \cdot dx.$$

$\nearrow$ C.D.F (differential is a PDF).

$= \cancel{E[E[l(y, g(x))]|x]}$

$= E\left[E\left[l(y, g(x)) \mid f_X(x)\right]\right]$

$$= \int_{\forall x} E[l(y, g(x))] \cdot f_x(x) \cdot dx$$

$$= \int_{\forall x} \left[\sum_{\forall y} [l(y, g(x))] \times f_{Y|X}(y/x) \right] \times f_x(x) \cdot dx$$

$\hookrightarrow$ ②

<u>digression</u>: $E(E(f_{Y|X}(y|x)))$

$= \int_{\forall x} E(f_{Y|X}(y|x)) \cdot f_x(x) \cdot dx$

$= \int_{\forall x} \left[\sum_{\forall y} y \cdot f_{Y|X}(y|x)\right] f_x(x)\, dx$

$= \int_{\forall x} \sum_{\forall y} y \cdot f_{Y/X}(y|x) f_x(x) \cdot dx$

$= \int_{\forall x} \sum_{\forall y} y \cdot f(x|y) \cdot f_Y(y) \cdot dx = \int_{\forall x} f(x|y) \cdot \sum y \cdot f_Y(y) \cdot dx$

$= \int_{\forall x} E(Y) \cdot f(x|y) \cdot dx$

$E(E(f_{Y|X}(y|x))) = E(Y).$

for binary classifican:

$$g \int_{x : g(x) = 1} \left[l(1, 1) \cdot f_{Y|X}(1/x) + l(-1, 1) \cdot f_{Y/X}(-1/x) \right] dx + \int_{x : g(x) = -1} \left[l(1, -1) \cdot f_{Y/X}(1/x) + l(-1, -1) \cdot f_{Y/X}(-1/x) \right] \cdot dx$$

The optimal solun will be $g^*(x) = \begin{cases} 1 & \text{if } l(-1, 1) \cdot f_{Y|X}(-1/x) < l(1, -1) \cdot f_{Y|X}(1/x) \\ -1 & \text{else.} \end{cases}$

(BAYES OPTIMAL).

Homework

Markov's inequality: $P(X \geq a) \leq \frac{E[X]}{a}$

Proof: let $1_E(x)$ be an indicator variable $1_E(x) = 1$ if $x = E$ else 0.

$1_{X \geq a} = 1$ if $X \geq a$ OR 0.

$\Rightarrow a I_{X \geq a} \leq X \Rightarrow E[a I_{X \geq a}] \leq E[X] \Rightarrow a E[I_{X \geq a}] \leq E[X]$

$\Rightarrow 1 \times P[X \geq a] + 0 \times P[X < a] \leq \frac{E(X)}{a}$

$\Rightarrow P[X \geq a] \leq \frac{E[X]}{a}$.

Chebyshev's inequality: $P(|X - E[X]| \geq a) \leq \frac{Var(X)}{a^2}$

Strong law of large nos:

The sample average converges almost surely to the expected value.

$\bar{X}_n \xrightarrow{a.s} \mu$ when $n \to \infty \Rightarrow Pr(\lim_{n \to \infty} \bar{X}_n = \mu) = 1$.

Weak law:

The sample average converges in probability towards the expected value.

$\bar{X}_n \to \mu$ when $n \to \infty$ for any +ve number ϵ

$\lim_{n \to \infty} Pr(|\bar{X}_n - \mu| > \epsilon) = 0$

30/07/14: (LECTURE - 2) dt: 30/7/14.

- We want approximate classifier that's close to the Bayesian Optimal classifier.

$E[l.(y, g(x))] = \frac{1}{m} \sum_{i=1}^{m} l(y_i, g(x_i))$ $\{Z_n\} \xrightarrow{P} E[X]$.

$\delta_n = P[|Z_n - E(X)| > \epsilon]$.

Degenerate distribution at the ~~end of the day~~ end when the sample size is high.

$\hat{g}_m = \min_g \frac{1}{m} \sum_{i=1}^{m} l(y_i, g(x_i))$ g^* is Bayesian Optimal.

(Empirical Risk min). (ERM)

g^* and $\hat{g}^*$ are closed in Expectation.

$\boxed{E[\hat{g}] = g^*}$ (How far is $\hat{g}$ from $\overset{*}{g}$ is the more imp. question).

ERM is statistically constrainted $\boxed{\{R[\hat{g}_m]\} \xrightarrow{P} R[g^*]}$ True risk of empirical data should converge in optimal risk.

$\hookrightarrow$ 1

$\hat{R}_{10}[\hat{g}_{10}] \leq \hat{R}_{10}[g^*]$ (For a dataset as large as having 10 example, a function $\hat{g}_{10}$ that gives the perfect solu^n (i.e minimizes the risk to the maximum extent) may do better than the "Bayesian optimal" itself].
(May not be the best generalization though].

* So, what model we choose is more important than the algorithms.

$$\underset{g\in G}{\operatorname{argmin}} \hat{R}_m[g] \Downarrow \quad \left\{ R[\hat{g}_m] \right\} \xrightarrow{p} R[\bar{g}]$$

$$\leq \text{ (less than equal to)}$$

$$\underset{g\in G}{\operatorname{argmin}} R[\bar{g}_m] \Leftarrow$$

$$g^* \to \underset{g}{\operatorname{argmin}} R[g] \leq \text{ (less than equal to).}$$

the terms involving true risk are not easy to compute. But we have "empirical risk" which is easier to compute. Can we express in terms of empirical risk?

$$0 \leq R[\hat{g}_m] - R[\bar{g}] \longrightarrow ②$$

($\hat{R}$ corresponds to empirical risk)

$$\leq \underbrace{\left[R[\hat{g}_m] - \hat{R}_m[\hat{g}_m] \right]}_{T} + \left[\hat{R}_m[\hat{g}_m] - R[\bar{g}] \right] - \hat{R}_m[\bar{g}] + \hat{R}_m[\bar{g}].$$

$\to$ 2(A)

(we just added & subtracted $\hat{R}_m[\hat{g}_m]$ & $\hat{R}_m[\bar{g}]$)

Since $\hat{R}_m[\bar{g}] - R[\bar{g}] \xrightarrow{p} 0$ (weak law) & $\hat{R}_m[\hat{g}_m] - \hat{R}_m[\bar{g}]$ follows equ^n (2) [i.e $\cancel{\geq} 0$], the sufficient cond^n for ② to be true will be if "T" is zero.

It looks like $R[\hat{g}_m] - \hat{R}_m[\hat{g}_m] \xrightarrow{p} 0$ but it may not be TRUE (why?? Homework) [Hint: Independence assumption in weak law].

Equ^n 2(A) can be rewritten as:

$$0 \leq \max_{g} \left[R[g] - \hat{R}_m[g] \right] + (\text{some term} \geq 0) + 0.$$

$$\cancel{R[g]} := \xrightarrow{} 0 \text{ (to be sufficient for 2).}$$

LECTURE - 3

dt: 1/8/14

Theorem: $\boxed{\lim_{m\to\infty} P\left[\max_{f\in\mathcal{F}}\left(R[f] - \hat{R}_m[f]\right) > \epsilon\right] = 0}$ ~~we want: $\lim_{m\to\infty} (R[f] - \hat{R}_m[f]) > \epsilon$~~

Weak law of large numbers:

$$\lim_{m\to\infty} P\left[\left|\underbrace{\frac{1}{m}\sum_{i=1}^{m} l(y_i, \hat{f}_m(x_i))}_{} - E[\underbrace{l(y, \hat{f}_m(x))}_{}]\right| > \epsilon\right] = 0$$

(Assuming IID.

Z_i

But Z_i may not follow IID since two random variables as f may depend on all x_is. So Z_is are not independent typically.

---

digression:

$\{f_n\} \to f$

$\{f_n(x_i)\} \to f(x_i)$ (point wise)

$|f_N(x) - f(x)| < \epsilon \;\forall n.$
OR $\max_x |f_N(x) - f(x)| < \epsilon.$ $\Big\}$ Uniform convergence criteria.

Revise

So: $\lim_{m\to\infty}\left[\max_{f\in\mathcal{F}}\left(R[f] - \hat{R}_m[f]\right) > \epsilon\right] = 0 \longrightarrow (2)$

(There is no "mod" so it's called one sided uniform convergence).

Illustration:

$\mathcal{F} = \{f_1 \cdots f_n\}$ [Finite].

for finite set (2) can be written as.

$$\lim_{m\to\infty} P\left[\bigcup_{f\in\mathcal{F}} [\underbrace{R[f] - \hat{R}_m[f]}_{} > \epsilon]\right] \le \sum_{i=1}^{m} P\left[R[f_i] - \hat{R}_m[f_i] > \epsilon\right]$$

$$\frac{1}{m}\sum_{j=1}^{m} Z_{ij} \equiv E[l(y, f(x))] - l(y_j, f_i(x_j))$$

& $E[Z_{ij}] = 0$. (How)??.

$$\sum_{i=1}^{n} P\left[\frac{1}{m}\sum_{j=1}^{m} Z_{ij} > \epsilon\right] = \sum_{i=1}^{n} P\left[\underbrace{e^{\frac{1}{m}\sum_{j=1}^{m} Z_{ij} > \epsilon}}_{+ve} > e^{\epsilon}\right] \text{ OR } = \sum_{i=1} P\left[e^{st\frac{1}{m}\sum_{j=1}^{m} Z_{ij} > \epsilon} > e^{st}\right]$$

may be -ve
(can't apply Markov Inequality)

$$\leq e^{-s\epsilon} \sum_{i=1}^{m} E\left[e^{s/m \sum_{j=1}^{m} z_{ij}}\right] = e^{-s\epsilon} \sum_{i=1}^{m-1} \prod_{j=1}^{m} E\left[e^{s/m z_{ij}}\right]$$

[Moment Generating Function of Gaussian is Gaussian."
Fourier Transform of]

$\boxed{\text{HOFFENDING INEQ:}}$ if $E(X) = 0$, $X \in [a, b]$ +ve

$$\underbrace{E\left[e^{sX}\right] \leq e^{\frac{s^2(b-a)^2}{8}}}_{\text{Nice bound.}} \rightarrow M(s) \quad N\left(0, \frac{(b-a)^2}{4}\right).$$

(Moment generating fn of a gaussian).

We can also have $E\left[e^{sX}\right] < e^{sb}$ (But not very nice).
not tighter ⟶

$$e^{sx} \leq \frac{b-x}{b-a} e^{sa} + \frac{x-a}{b-a} e^{sb}.$$

$$E(e^{sX}) \leq \left(\frac{b}{b-a}\right) e^{sa} + \left(\frac{-a}{b-a}\right) e^{sb} = e^{h(z)}, \quad = -\theta z + \ln(1-\theta+\theta e^{z}).$$

$1-\theta \qquad \theta$

$$(1-\theta)e^{sa} + \theta e^{sb}.$$

if $z = s(b-a)$

$$= (1-\theta)\left[1 + sa + \frac{s^2a^2}{2!} + \frac{s^3a^3}{3!} \cdots\right] + \theta\left(e^{sb}\right.$$

$h(0) = 0$

$h'(0) = 0$

$$h''(z) = \frac{d}{dz} \frac{1}{1-\theta+\theta e^{z}} \times \theta \cdot e^{z}.$$

$$= e^{sa} \times \theta\left[s(b-a) + \frac{s^2(b-a)^2}{2!} \cdots\right]$$

$$= \frac{1}{1-\theta+\theta e^{z}} \times \theta e^{z} + \theta e^{z} \times \frac{-1}{(1-\theta+\theta e^{z})^2} \times \theta e^{z}.$$

$$= \frac{\theta e^{z}(1-\theta)}{(1-\theta+\theta e^{z})^2} \leq \frac{1}{4}$$

$$h(z) \leq \frac{z^2}{8}$$

(Revise Taylor series).

$$\frac{a+b}{2} \leq \sqrt{ab}.$$

Lecture : 4 dt: 06/08/14

Recap

① Our goal has been bringing the risk as close as to the optimal risk.

Statistical constraint: $\{R[\hat{f}_m]\} \xrightarrow{P} R[\bar{f}] \Leftrightarrow \lim_{m\to\infty} P[\max_{f\in\mathcal{F}} R[f] - \hat{R}_m[f]] = 0$

Bayesian constraint: $\{R[\hat{f}_m]\} \to R[f^*]$

(IMPORTANT POINT STARTING POINT)

② We tried taking $\mathcal{F}$ from a finite set. $\mathcal{D} \to$ i.i.d.

$$\lim_{m\to\infty} P\left[\max_{f\in\mathcal{F}} R[f] - \hat{R}_m[f] > \epsilon\right] = 0 \longrightarrow ①$$

(Refer to previous lecture note)

$$\le e^{-s\epsilon} \sum_{i=1}^{n} \prod_{j=1}^{m} E\left[e^{s/m\, z_{ij}}\right]$$

(we applied Markov inequ)
(we assumed i.i.d for absorbing the prod)

$$\le e^{-s\epsilon} \times n \times e^{s^2/m \cdot \frac{(b-a)^2}{8}} \quad \forall A > 0.$$

(we ~~cancelled~~ got rid of the product)

$$\boxed{\text{lets substitute } s = \frac{4m\epsilon}{(b-a)^2}}$$

(objective: To get tightest bound)

$$\le n\, e^{\frac{-2m\epsilon^2}{(b-a)^2}}$$

(Using Hoffending inequality).

comment

When $n \to \infty$, the above term will be zero. so equ ① will be satisfied [∵ lower bound = 0 upper bound = 0 (sandwiching)]

$$P\left[R[\hat{f}_m] - \hat{R}_m[\hat{f}_m] > \epsilon\right] \le n e^{\frac{-2m\epsilon^2}{(b-a)^2}} \longrightarrow ②$$

$\to \delta$ (say) $\to$ (2a)

$$\Rightarrow P\left[R[\hat{f}_m] - \hat{R}_m[\hat{f}_m] \le \epsilon\right] \ge 1-\delta \longrightarrow 2(b).$$

with probability $1-\delta$, we have $\forall f \in \mathcal{F}$

$$R[\hat{f}_m] \le \hat{R}_m[\hat{f}_m] + \epsilon \longrightarrow ③$$

Generalizing ③ by saying that if the "max" deviaⁿ is less than ϵ then each deviation will also be less than ϵ $\forall f \in \mathcal{F}$

$$R[f] \le \hat{R}_m[f] + (b-a)\sqrt{\frac{\log n/\delta}{2m}} \to ④$$

(substituted ϵ by δ from 2a).

Equⁿ ④ is Known as LEARNING BOUND. (Remember it is a probabilistic bound).

The LEARNING BOUND may be satisfied for any kind of Risk minimiz? but we have to go back to the "starting point" and check if that holds true.

Probably Approximately Correct.

$\mathcal{F}$ is (agnostic) PAC-learnable if for an $\epsilon > 0$; $\delta \in (0,1)$, D, size(f)

$x \in \mathbb{R}^D$ $\quad m \geq poly\left(\frac{1}{\epsilon}, \frac{1}{\delta}, D, size(f)\right) \rightarrow P\left[R[\hat{f}_m] - \hat{R}_m[\hat{f}_m] \leq \epsilon\right] \geq 1-\delta$

suppose we take $m = \frac{(b-a)}{2\epsilon^2} \log n/\delta$ (polynomial of the above described components).

Equ^n 2(b) will be true for our 'm' and also for $m \geq \frac{(b-a)}{2\epsilon^2} \log n/\delta$

$\boxed{\text{So, finite } \mathcal{F} \text{ is always PAC-learnable.}}$

Example:

Let $\overset{\in(0,1)}{x_1}, \overset{\in(0,1)}{x_2} \dots \overset{\in(0,1)}{x_d}$ are properties of an object. and ideally if a decision is positive if $x_1 \wedge \bar{x}_2 \wedge x_9 = f$ is 1.

so f is our f^* (optimal), Now let's consider some training example. $\boxed{|f| = 3^d}$ (for d ~~examples~~ terms) ⇄ same compliment skipping

x_1	x_2	x_3	x_4	
1	0	1	1	→ +ve
0	0	1	1	→ −ve

what ERM can give us that we can exactly fit the training data to ~~come~~ arrive at a "conjunction" which will not fail for training data. (or at least the loss will be minimum).

For d ~~examples~~, our maximum loss can be 'd' and minimum loss can be 0. $(a=0, b=d)$.

$$m \geq \frac{d}{2\epsilon^2}\left(d \log 3 + \log \frac{1}{\delta}\right) \quad \text{(PAC learnable)}.$$

[PAC-learnability is a more general case and it doesn't always consider ERM]

Lecture-5 dt: 07/08/14

Recap: ERMs for finite function classes are always statistically constraints. → ①

② $P\left[\max_{f \in \mathcal{F}} R[f] - \hat{R}_m[f] > \epsilon\right] \leq n e^{-2m\epsilon^2/(b-a)^2}$ → ②.

③ ~~$m \geq (b-a)^2 \sqrt{\frac{\log n/\delta}{2\epsilon^2}}$~~ $m \geq \frac{(b-a)^2}{2\epsilon^2} \log n/\delta$ → ③ [correct the previous lecture note].

finite function classes are PAC learnable. "log" is more helpful (why??).

Learning Rate $\epsilon = O(1/\sqrt{m})$ [slower learning rate]

A rich function class $\mathcal{F}$ should achieve $\boxed{\hat{R}_m[\hat{f}_m] = 0}$

From discussion of lecture-2:

$$P\left[R[\hat{f}_m] - \hat{R}_m[\hat{f}_m] > \epsilon\right]$$

$$= P\left[R[\hat{f}_m] > \epsilon\right] \leq \frac{E[R[\hat{f}_m]]}{\epsilon}$$

(Using this method I can't derive easily the rate of learnability).

$$\leq \sum_{i : R[f_i] > \epsilon} P\left[\hat{R}_m[f_i] = 0\right]$$

(Total probability rule)

$$\leq \sum_{i : R[f_i] > \epsilon} (1-\epsilon)^m \quad \text{[for 0-1 loss]}.$$

Revise

$R[f_1] > \epsilon \quad R[f_{10}] > \epsilon$

$f_1 \ldots f_{10}$

[for all such f_i which true risk is greater than ϵ, are candidate $\hat{f}_m$, so the condition $\hat{R}_m[\hat{f}_m]$ is zero.]

$\leq n \cdot (1-\epsilon)^m$ (will go to 0 if $m \to \infty$ so statistical consistency is satisfied).

Using the bounding condition $1 - \epsilon \leq e^{-\epsilon}$ (use convex property of $e^{-\epsilon}$) for any ϵ

$$\leq n \times e^{-m\epsilon} \longrightarrow ④.$$

Learning rate $\epsilon = O\left(\frac{1}{m}\right)$ (FASTER)

[For non 0,1 loss, there will be some a, b terms in the expressⁿ. Prove that.]

Book Ref: Foundⁿ of ML by "Mohri".

Lecture-6 dt: 13-8-14

Arbitrary function classes (need not be finite).

$$P\left[\max_{f\in\mathcal{F}} R[f]-\hat{R}_m[f]>\epsilon\right] \leq \_\_\_\_ ?$$

In case of finite $\mathcal{F}$ we converted to union bounding.

$$\leq \sum_{f\in\mathcal{F}} P\left[R[f]-\hat{R}_m[f]>\epsilon\right]$$

(Best possible bound without additional assumptions Chernoff Bound ...).

2^m equivalence classes of functions in $\mathcal{F}$.

① Can we have a bound in case of arbitrary (non-finite) $\mathcal{F}$? What is the intuition?

→ There are many functions which will correspond to the same empirical risk (in most cases the "loss" doesn't change). The maximum of such equivalence class ~~may have~~ in $\mathcal{F}$ ~~2m functions~~ will be in the order of 2^m.

$$\max_{f \text{ in } \mathcal{F}} R[f] = R[f']$$

$$\therefore\ P\left[\max_{f\in\mathcal{F}} R[f]-\hat{R}_m[f]>\epsilon\right] \leq P\left[\max_{\mathcal{F}\text{ in } f} R[f']-\hat{R}_m[f]>\epsilon\right]$$

Considering the above equation and applying the notion of equivalence classes (2^m) and working out with chernoff-bounding & Hoffending bound we can still be able to prove that the term will converge to some value (may not be 0) irrespective of the size of $\mathcal{F}$ ($\mathcal{F}$ may well be ∞ set).

② $$P\left[\max_{f\in\mathcal{F}} R[f]-\hat{R}_m[f]>\epsilon\right]$$

↳ Random variable $g(v_1 \cdots v_m)$ $v_i=(x_i,y_i)$

↳ $v_1, v_2 \cdots v_m$ are iid

↳ bounded difference property.

One dimentional variation is bounded by a constant

$$|g(v_1,v_2\cdots v_m)-g(v_1,v_2\cdots v_k'\cdots v_m)| \leq \frac{b-a}{m} \text{ (Prove)}.$$

(Recall) (finite case)

$$\sum_{f\in\mathcal{F}} P[R[f]-\hat{R}_m[f]>\epsilon] \quad \underset{\forall f}{\sum}\sum_{i=1}^{m} \frac{E(\ell(y,f(x))-\ell(y_i,f(x_i))}{m}$$

$$\underset{\forall f}{\sum}\sum_{i=1}^{m} z_i \qquad E(z_i)=0$$

In finite case 'g' was nothing but a sum function with variance

$$\operatorname{var}\sum_{i=1}^{m} z_i = \sum_{i=1}^{m}\operatorname{var}(z_i) \leq \left(\frac{b-a}{m}\right)^2/4$$

$\Rightarrow$ Without applying chernoff Bound, let's go ahead..

$$P[g>\epsilon] = P[e^{sg}>e^{s\epsilon}] \leq e^{-s\epsilon}E[e^{sg}] \text{ (Markov).}$$

$$= e^{-s\epsilon}E[E[e^{sg}/U_1,U_2\ldots U_{m-1}]] \longrightarrow (1)$$

The introduction of conditional probability was made to enable applicⁿ of Hoffending bound but to apply that $E[g]$ should be "zero" (which is not in our case).

Trick: $\max_{f\in\mathcal{F}}(\cdot) \geq (\cdot) \Rightarrow E[\max_{f\in\mathcal{F}}(\cdot)] \geq E(\cdot) \Rightarrow E[\max_{f\in\mathcal{F}}(\cdot)] \geq \max_{f\in\mathcal{F}} E(\cdot)$

Use: notion $E[E(g/U_1\ldots U_{m-1})/U_1\ldots U_{m-2}] = E(g)$

$$= e^{-s\epsilon}E\left[E\left[e^{s[(g/U_1,\ldots U_{m-1})-E[g/U_1,\ldots U_{m-1}]]+sE[g/U_1,\ldots U_{m-1}]}\right]\right]$$

(now we can apply Hoffending bound.

$$\leq e^{-s\epsilon}E\left[e^{sE(g/U_1\ldots U_{m-1})}\cdot e^{s^2(b-a)^2/8m^2}\right]$$

If we keep on adding and subtracting like this for $g/U_1,\ldots U_{m-2}$, $g/U_1,\ldots U_{m-3}$ .... we will have $E(g)$ in the final step.

[Prove that these terms are also satisfying BD]

→ Repeated 'm' number of times,

$$\leq e^{-s\epsilon}E\left[e^{sE(g/U_1\ldots U_{m-1})}\cdot e^{s^2(b-a)^2/8m^2}\right]$$

$$\leq e^{-s\epsilon}E\left[e^{sE[g]}\cdot e^{s^2(b-a)^2/8m^2\times m}\right] \leq e^{-s\epsilon}e^{\frac{s^2(b-a)^2}{8m}}\cdot e^{sE(g)}$$

Following similar discussion as lecture-4,

$$P[g>\epsilon] \leq e^{-2m[\epsilon-E(g)]^2/(b-a)^2}$$

McDiarmid Inequality: $\boxed{P[g-E(g)>\epsilon] \leq e^{-2\epsilon^2/\sum_i c_i^2} \;\&\; P[E(g)-g>\epsilon] \leq e^{-2\epsilon^2/\sum_i c_i^2}}$

LECTURE - 7   dt: 14/8/14.

Recall: $P\left[\underline{\max_{f\in\mathcal{F}} R[f]-\hat{R}_m[f]}>\epsilon\right]$

$\hookrightarrow g \to c_i = \frac{b-a}{m}$

McDiarmid g — independent rvs as per — bounded dif.

$$P[|g - E[g]| > \epsilon] \leq e^{-2\epsilon^2/\sum_i c_i^2} \longrightarrow (1).$$

Refer Proof of the "Bounded difference inequality" from Berkeley.edu/~bartlet

From previous lecture, we had:

$$P[g > \epsilon] \leq e^{-2m[\epsilon - E(g)]^2/(b-a)^2} = \delta \text{ (say)}$$

We can say, with probability $1-\delta$

$$g \leq (b-a)\sqrt{\frac{\log 1/\delta}{2m}} + E[g]$$

$$\Rightarrow R[f] \leq \hat{R}_m[f] + E[g] + (b-a)\sqrt{\frac{\log 1/\delta}{2m}} \quad \forall f \in \mathcal{F}$$

$\hookrightarrow (2)$ not finite.

(Recall) For finite case we had

$$R[f] \leq \hat{R}_m[f] + (b-a)\sqrt{\frac{\log |\mathcal{F}|/\delta}{2m}} \quad \forall f \in \mathcal{F} \text{ (finite)}$$

$\hookrightarrow (2a)$

Speculation: This bound may be tighter than (2a) because (intuition) (2a) follows union bound (which may be unnecessary).

We will evaluate $E[g]$ later. But $E[g]$ should decay at least faster than $O(1/\sqrt{n})$ to make it better than (2a). So, choosing a function class $\mathcal{F}$ that helps $E[g]$ decay faster is necessary.

Now let's have:

$$\Omega(\mathcal{F}) \equiv E\left[\max_{f\in\mathcal{F}} R[f] - \hat{R}_m[f]\right] \geq 0$$

(Recall $E(\max(\cdot)) \geq \max[E(\cdot)]$)

[If: $\Omega(\mathcal{F}_1) \geq \Omega(\mathcal{F}_2)$ then $\mathcal{F}_1 \supseteq \mathcal{F}_2$ (monotonic)]

~~$\Omega(\mathcal{F}_1 \cup \mathcal{F}_2) = \max(\Omega(\mathcal{F}_1), \Omega(\mathcal{F}_2))$~~

~~$\leq \Omega(\mathcal{F}_1) + \Omega(\mathcal{F}_2)$ (subadditive)~~]

$$\Omega(\mathcal{F}) = E\left[\max_{f\in\mathcal{F}} \underbrace{R[f] - \hat{R}_m[f]}\right]$$

lets say $R[f]$ is coming from some other set of data.

$$\cancel{E\left[\max_{f\in\mathcal{F}} R'_n[f] - \hat{R}_m[f]\right]} \qquad \cancel{R'_n[f] = \sum_{i=1}^{n} l(y_i, f(x_i))}$$

$$= E\left[\max_{f\in\mathcal{F}} \; E_{x',y'}\left[\frac{1}{m}\sum_{\forall x',y' \; \forall i=1 \text{ to } m} l(y'_i, f(x'_i)) - \frac{1}{m}\sum_{i=1}^{m} l(y_i, f(x_i))\right]\right]$$

$$\leq E\left[\underbrace{\max_{f\in\mathcal{F}} R'_m[f] - \hat{R}_m[f]}_{\text{max. discrepancy of } \mathcal{F} \Rightarrow D(\mathcal{F})}\right]$$

discrepancy for at-least one function

Equ^n (2) becomes:

$$E[D(f)] \leq D(\mathcal{F}) + (b-a)\sqrt{\frac{\log 1/\delta}{2m}} \rightarrow \text{(2b)}.$$

so with prob $1-\delta$:

$$R[f] \leq \hat{R}_m[f] + D(\mathcal{F}) + 2(b-a)\sqrt{\frac{\log 1/\delta}{2m}} \rightarrow \text{(3)}$$

Guaranteed Risk.

intuitively,

A function class is supposed to be good if $D(\mathcal{F})$ is minimum. Even if we have at least one function in the class of $\mathcal{F}$ which maximizes $D(\mathcal{F})$ to a large quantity, the function class will still be bad.

$$E\left[\max_{f\in\mathcal{F}} \hat{R}_m[f] - \hat{R}_m[f]\right] \qquad \delta = \{1, -1\} \text{ (say)}.$$

$$= E\left[\max_{f\in\mathcal{F}} \frac{1}{m}\left(\sum_{i=1}^{m} \delta_i [l(y'_i, f(x'_i)) - l(y_i, f(x_i))]\right)\right]$$

$$\leq E\,E_\delta\left[\max_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} \delta[l(y'_i, f(x'_i))]\right] + E\,E_\delta\left[\max_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} -\delta[l(y_i, f(x'_i))]\right]$$

(Assuming δ is uniformly distributed

$$= 2E\,E_\delta\left[\max_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} \varepsilon_i\, l(y_i, f(x_i))\right]$$

(Radmacher's complexity) $\mathcal{R}_m(\mathcal{F})$

Now applying total law of Expect^n on $\hat{R}_m$

$$\underbrace{2E_{\sigma}\left[\max_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i \, l(y_i, f(x_i))\right]}$$

$$= E\left\{E_{\sigma}\left[\max_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i \, l(y_i, f(x_i)) \Big/ \left[(x_1, y_1), (x_2, y_2) \ldots (x_m, y_m)\right]\right]\right\}$$

(Computing conditional R_m is easy.)

Now applying chernoff bounds, union bounds etc..

for a probability $1-\delta$:

$$R[f] \leq \hat{R}_m[f] + 2R_m[\mathcal{F}] + 3(b-a)\sqrt{\frac{\log 1/\delta}{2m}} \quad \forall f \in \mathcal{F}$$

$\hookrightarrow$ (4)

Lecture - 8

Dt: 20th Aug

Recall:

$\mathcal{F}$ is finite

$$R[f] \leq \hat{R}_m[f] + (b-a)\sqrt{\frac{\log |\mathcal{F}|/\delta}{2m}} \quad \forall f \in \mathcal{F} \longrightarrow (1)$$

$\mathcal{F}$

$$\hookrightarrow R[f] \leq \hat{R}_m[f] + \hat{D}_m[\mathcal{F}] + (b-a)\left(\sqrt{\frac{\log 2/\delta}{2m}} + \sqrt{\frac{\log 2/\delta}{m}}\right) \forall f \in \mathcal{F} \longrightarrow (2)$$

$$\downarrow$$

$$\max_{f \in \mathcal{F}} \hat{R}'_m[f] - \hat{R}_m[f]$$

$$\Rightarrow R[f] \leq \hat{R}_m[f] + 2\hat{\mathbb{R}}_m[\mathcal{F}] + 3(b-a)\sqrt{\frac{\log 2/\delta}{2m}} \quad \forall f \in \mathcal{F} \longrightarrow (3)$$

$$\downarrow$$

$$E_\sigma\left[\max_{f \in \mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i l(y_i, f(x_i))\right]$$ (Rademacher's function).

$$\hat{\mathbb{R}}_m[\mathcal{F}] = E\left[\max_{f \in \mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i f(z_i)\right].$$

$$= E\left[\max_{f \in \mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} \sigma_i (f(z_i) / \forall z_i)\right]$$

So $\hat{\mathbb{R}}$ is, as per equation (3) not only a function of function classes, it is a combination that contains the "loss" information. Hence $\mathbb{R}(f)$ is sometimes replaced by $\mathbb{R}(L)$ or Rademacher function for "loss class".

Example: $\mathbb{R}$ of linear classifiers.

lets say $l \rightarrow$ zero one loss

$$f \rightarrow \{ f \mid \exists w \in \mathbb{R}^N \ni f(x) = \text{sign}(w^T x) \ \forall x \in \mathbb{R}^N \}$$

$$\hat{\mathbb{R}}(\mathcal{F}) = E\left[\max_{w \in \mathbb{R}^N} \frac{1}{m}\sum_i \sigma_i \underbrace{\frac{(1 - y_i \text{sign}(w^T x_i))}{2}}_{(0,1 \text{ loss func}^n)}\right]$$

$$\Rightarrow 2\hat{\mathbb{R}}(\mathcal{F}) = E\left[\max_{\underbrace{w \in \mathbb{R}^m}_{z \in \mathcal{Z} \subset \{-1,1\}^m}} \frac{1}{m}\sum_i \sigma_i \underbrace{y_i \text{sign}(w^T x_i)}_{z_i}\right] \longrightarrow (4).$$

<u>Trick:</u> Using Jensen's inequality i.e $f(E(x)) \leq E(f(x))$,
we can transform $E(x) \leq \frac{1}{s}\log(E[e^{sx}])$.

Equ^n (4) can now be transformed to:

$$2\hat{R}(\mathcal{F}) \leq \frac{1}{s}\ln\left(E\left[e^{s \max_{z_i} \frac{1}{m}\sum_i \delta_i z_i}\right]\right).$$

$$= \frac{1}{s}\ln\left(\max_{z_i} E\left[e^{s\frac{1}{m}\sum_i \delta_i z_i}\right]\right)$$

($e^{\max(x)} = \max e^{(x)}$ monotonic).

$$= \frac{1}{s}\ln\left(\sum_{z_i} E\left[e^{s\frac{1}{m}\sum_i \delta_i z_i}\right]\right)$$

e^x is +ve so $\max(e^x) <$ sum(e^x).

$$= \frac{1}{s}\ln\left(\sum_{z_i}\prod_{i=1}^{m} E\left[e^{s/m\, \delta_i z_i}\right]\right)$$

Hoffeding → [for zero-one loss $\frac{b-a}{m} = 2/m$]

$$= \frac{1}{s}\ln\left(\sum_{z\in Z} e^{s^2/2m}\right) = \frac{\ln(|Z|)}{s} + s/2m \longrightarrow 5.$$

differencing w.r.t. s and equating to 0 (for finding the tightest bound).

$$\boxed{2\hat{R}(f) \leq 2\sqrt{\frac{\log|Z|}{2M}}} \longrightarrow 6.$$

$|Z|$ is reffered to as growth functions of 'Z'. (Takes a value of m and returns a number)

$|Z_m|$ is upper bounded by 2^M which when replaced in 6 gives a positive value $\left(2\sqrt{\log 2/2}\right)$ (not bad).

Now can we have better bound than saying $|Z_m| \leq 2^m$? (n is the dimension)

one bound is $|Z_m| \leq {}^mC_n$ ~~(n+1) the number of points the plane passes through~~).

$|Z_m| \leq 2^{n+1}\, {}^mC_n$ (our speculation)

But according to theory, classical bound is

$$|Z_m| \leq \sum_{i=0}^{n+1} {}^mC_i \quad \text{(Prove)}$$

Lecture - 9

dt: 22 Aug.

We were trying to give bound to τ_m or $\Pi_f(m)$, the growth funcⁿ of linear classifiers.

$$\Pi_f(m) \begin{cases} \leq 2^{n+1}\, {}^mC_n \text{ (in class)} \leq \left(\frac{me}{n}\right)^n 2^{n+1} \rightarrow ① \\ \leq \sum_{i=0}^{d} {}^mC_i \text{ (classical theory)} \approx \left(\frac{m}{n+1}\right)^{n+1} \rightarrow ② \end{cases}$$

VC-dimensionⁿ. d is V.C of $\mathcal{F}$

(we used $\left(\frac{m}{n}\right)^n \leq {}^mC_n \leq \left(\frac{me}{n}\right)^n$)

In some cases ① may be tighter than ② but considering all possible value of n, m & $\mathcal{F}$, ② decays faster.

V.C is the d where there are 'd' points that can be shattered by $\mathcal{F}$.

Proof of ②:

Π_f is the no. of rows

	x_1	x_2	. . .	x_n
f_1	+	−		+
:				
f_n	−	+		−

→ distinct funcⁿ. (at least one value should be different from other row).

↓ applying transformⁿ. (Let us assume that we can have another function class $\mathcal{F}^*$ by changing the sign of at least one value in each row such that two rows don't become identical.

	x_1	. . .	x_n
f_1	−	−	+
:			
f_n	−	+	−

[By constructing examples we can say that the points which are shattered in table-1 may not be shattered in table-2 (VC dimension changes). But for those points which are shattered in 2 should be shattered in 1.]

1 ⟶ 2 (introducⁿ of more shattered points)

2 ⟶ 1 (Already retain shattered points)

[Reason: If we change the shattered points in 1 as a result of transformⁿ, we will end up creating duplicates, so that change is not allowed, hence shattered points don't change.]

so; $\boxed{vc(\mathcal{F}) \leq vc(\mathcal{F}^*)}$. (Digression)

we had $\hat{\mathcal{R}}_m(\mathcal{L}) \leq \sqrt{\frac{2 \log \Pi_g(m)}{m}}$ → PAC learnable.

so; $R[f] \leq \hat{R}_m[f] + 2\sqrt{\frac{2n \log(m/n\, e)}{m}}$ (substituting 1).

Our linear classifier becomes:

$\min_{w \in \mathbb{R}^n} \frac{1}{m} \sum_i 1\{ y_i w^T x_i \neq y_i \}$ (computationally hard to find out the combinations)

To make the $\mathcal{F}$ close to Bayes optimal. problems with linear classifiers

① we have to change the loss fun^n (what about non-linear functions?) if B.O is non linear.

② To account for that if we replace x by $\phi(x)$ which will, say, ~~give~~ take us closer to B.O. Are we doing better?

② There is an effect of 'n' on the complexity of taking f closer to Bayes optimal. ("Curse of dimensionality")

[Gausian complexity; if we replace σ_i, Rademacher's variable by a standard gausian variable (with mean 0 & stdev 1)]

we will have $\mathbb{E}_{g_i}\left[\max_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} g_i f(z_i) \right]$

choice of loss fun^n:

Binary Classifiers: (Refer 0-1 loss, hinge loss, squared hinge loss, logistic loss).

[Prove $\log(1 + e^{-yf(x)})$ is convex in w]. (convex not diff.) convex (diff).

Regression: loss → $|y - f(x_i)|$ squared loss: $[y - f(x_i)]^2$

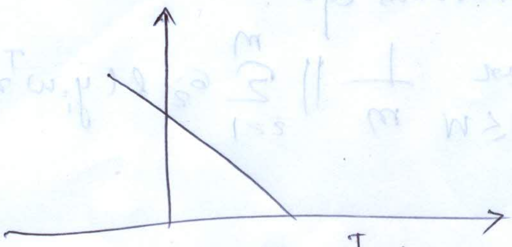
A loss function is good if it is convex, differentiable, ~~and~~ also address sparsity (i.e, we can skip the examples for which loss is zero if we have prior knowledge about such examples) (↳ and we still get the same solu^n).

IMP: Most of the losses (except 0-1) are not bounded. But the data can help us put a bound on $\|w\|$ which will help us bound the loss function.

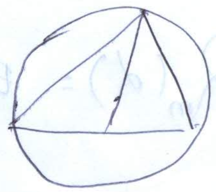
$w^T x \leq \|w\| \|x\| \leq \|w\| R$

Lecture - 10 dt: 26/8/14. (Credit: Anoop).

Hinge loss:



$y_i w^T x_i$



R → Max bound on.

V.C dimension of thick/fat classifier (for 2D case):

$$VC = \begin{cases} 1 & \text{if } M > 2R \\ 2 & \text{if } \frac{3R}{2} < M \leq 2R \\ 3 & \text{if } M \leq \frac{3R}{2} \end{cases}$$

where M → margin $\left(\frac{2}{\|w\|}\right)$.

when $\|w\|$ increases, $M \to 0$, we will have a large V.C dimension.

* VC dimension can be controlled.

For 3D case:

$$VC \leq \min\left(\left\lceil \frac{4R^2}{M^2} \right\rceil, n\right) + 1 \rightarrow \text{for fat classifiers.}$$

Fat classifiers give control over VC dimension.

(Note: Fat classifiers ≠ classifiers with Hinge loss).

VC dimensions defined only for binary classes.

Radamacher's Complexity:

$$F = \left\{ f \mid \exists \text{ a } w \in \mathbb{R}^N \rightarrow f(x) = w^T x \ \forall x, \|w\| \leq W \right\}$$

$$\hat{R}_m(\alpha) = E\left[\max_{\|w\| \leq W} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \max(0, 1 - y_i w^T x_i)\right]$$

Now, by Holder/Schwarz inequality:

$$\max_{\|w\| \leq W} \frac{1}{m} \sum_i \sigma_i w^T x_i \leq W \left\| \frac{1}{m} \sum_i \sigma_i x_i \right\|_2$$

A general form of the previous eqⁿ:

$$\hat{R}_m(\alpha) = E\left[\max_{\|w\| \le W} \frac{1}{m} \left\| \sum_{i=1}^{m} \sigma_i \, l(y_i, w^T x_i) \right\|_2 \right] \rightarrow (1).$$

<u>Contraction lemma: Proof:</u>

$l \rightarrow$ Lipschitz continuity

$$|l(x) - l(y)| \le L \|x - y\| \quad \forall x, y.$$

(See Meir lemma: meir2003.pdf).

A convex function on any compact subset of "interior of domain" is Lipschitz continuous.

We will apply the theorem to Rademacher average of "Hinge loss" like loss function.

Lecture: 11 dt: 3/9/14

$$\hat{\mathcal{R}}_m(\ell) \leq \hat{\mathcal{R}}_m(\mathcal{F}) = E\left[\max_{\|w\|<W} \frac{1}{m}\sum_i \sigma_i y_i w^T x_i\right]$$ ($\ell \to$ hinge loss; $\mathcal{F} = \{$lin functions $\|w\| \leq W\}$.)

$$w^T\left(\frac{1}{m}\sum_i \sigma_i y_i x_i\right)$$

$$\leq \frac{W}{m} E\left[\left\|\sum_i \sigma_i y_i x_i\right\|\right]$$ (Applying Lipschitz continuity: contraction lemma)

$$\leq \frac{W}{m} E\left[\sqrt{\left\|\sum_i \sigma_i y_i x_i\right\|^2}\right]$$

Applying Jensen's inequality (since square root is convex. $E(f(x)) \leq f(E[x])$)

$$\leq \frac{W}{m}\sqrt{E\left[\left\|\sum_i \sigma_i y_i x_i\right\|^2\right]}$$

$$= \frac{W}{m}\sqrt{E\left[\sum_i\sum_j \sigma_i\sigma_j y_i y_j x_i x_j\right]}$$

$$= \frac{W}{m}\sqrt{\sum_i \|x_i\|^2}$$

$$\leq \frac{WR}{\sqrt{m}}$$

$$R[f] \leq \hat{R}_m[f] + \frac{2WR}{\sqrt{m}} + 3\sqrt{\frac{\log 2/\delta}{m}} \quad \forall f \in \mathcal{F} \longrightarrow (1)$$

Extension to non-linear classifier:

[Reason: Bayes optimal for an application need not be linear].

$f(x) = w^T\phi(x)$ $\qquad \phi: x \to \mathbb{R}^N$ $\qquad$ Example: $\phi(x) = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \\ x_1 x_2 \end{bmatrix}$

$\phi(x)$ is not only a function giving rise to vectors with finite dimensions. $\phi(x)$ can produce a function that can take x and produce values. (may facilitate ~~taking~~ vectors with infinite dimension).

(Revise 'vectors', 'vector space', inner products)

(we can replace $\phi(x)$ in the above derivation (eqn 1) at least for $\phi(x)$ to be belonging to Euclidian space). More discussion to happen on this in the next lecture.

~~The~~ $$\min_{w \in \mathbb{R}^N} \frac{1}{m} \sum_i l(y_i, w^T \phi(x_i)) \qquad \text{(Projected Gradient descent)}$$

$$\text{s.t: } \|w\| \leq W$$

[Terminologies: function class $\equiv$ Model $w \equiv$ model parameters
$W \equiv$ hyper parameter]

When we change the hyper parameters, the model changes.

Now, how to automatically select W. → find smallest W where $\hat{R}_m = 0$.

→ $$\min_{w \in \hat{\mathbb{R}}, W} \sum_i l(\ldots) + \frac{2WR}{\sqrt{m}}$$

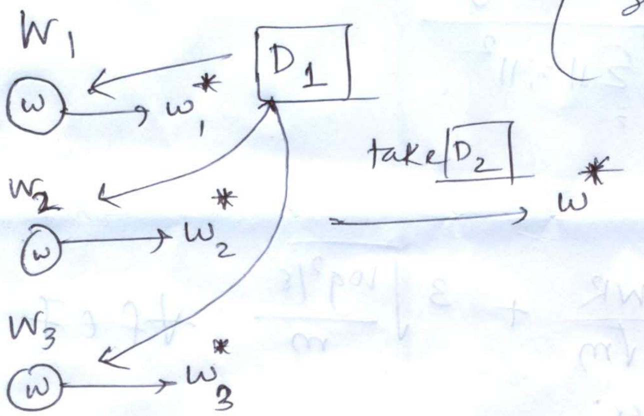
$$\|w\| \leq W \ldots$$

OR

↓

Take a separate training set and do ERM for finding out W.
(called validation).

(Surprisingly, this works better in practice).



Lecture: 12 dt: 05/09/14

(I) Two ways to choose W:

① Structural Risk Minimiz^n: $\min_{w \in \mathbb{R}^N, W} \sum_i \ell(\cdots) + \frac{2WR}{\sqrt{m}}$ $\|w\| \leq W$.

② ERM: validation using different datasets.

(II) 'ϕ'
- → Statistical consistency $\|\phi(x)\| \leq R$ (Bounded).
- → Bayes consistency ("Universal").
- → X should be generic (any kind of input should be accepted).
- → Computationally efficient. (also non linear). (A polynomial function with degree d with dimension D can grow exponentially in terms of complexity when the input is large).

Kernels

Kernels:

Let consider X, a set (arbitrary set).

ϕ → $\phi(x)$ → $f: X \to \mathbb{R}$

↳ (Assume continuity)

$\mathcal{H}$ → (should be some generalization of Euclidian spaces to deal with infinite dimension).

We seek that $\phi(x)$ in the space of H should be allowed to have the following properties:

+, $\underset{\text{(scaling)}}{\cdot}$ → $\phi(x) \subseteq \underset{\text{Linear Hull (combinations)}}{\mathrm{LIN}}(\phi(x))$

Inner product $\langle\ ,\ \rangle$ → ⓐ $\langle x, x\rangle \geq 0$, $\langle x, x\rangle = 0 \Leftrightarrow x = 0$ } Non-neg.

ⓑ symmetry: $\langle x, y\rangle = \langle y, x\rangle$

ⓒ $\langle \alpha x + \beta y, z\rangle = \alpha\langle x, z\rangle + \beta\langle y, z\rangle$ (linearity).

$\phi(x)$ itself is a function that can take parameters.

(e.g. $\phi(x)(z) = 3$)

$\phi(x): X \to \mathbb{R}$ (captures similarity??). → ①.

(The "math" way to capture similarity is to take the inner product (may be with some normaliz^n)).

$\langle \phi(x), \phi(z)\rangle \to \phi(x)(z)$. OR

$f = \sum_i \alpha_i \phi(x_i)$ $g = \sum_j \beta_j \phi(x_j)$.

$\langle f, g\rangle = \sum_i \sum_j \alpha_i \beta_j \phi(x_i)(x_j)$.

Example: ① $\phi(x) = x \rightarrow \phi(x)^T \phi(z) = x^T z$.

② $\phi(x)(z) = x^T z \Rightarrow \langle \phi(x), \phi(z) \rangle = x^T z$.

similar representations

but in the second case, ϕ is a function returning a function. Both representations are similar in terms of dot product.

③ $\phi(x)(z) = (x^T z + 1)^d$ (say), will be equivalent to $\langle \phi(x), \phi(z) \rangle$

for $\phi(x) = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \\ \sqrt{2} x_1 x_2 \end{bmatrix}$ $\left\{ \langle \phi(x)^T \phi(z) \rangle = (1 + x^T z)^2 \right\}$ $d = 2$.

④ $\phi(x)(z) = e^{x^T z} = \langle \phi(x), \phi(z) \rangle$ for $\phi(x) = e$ (variant of e^{x^T})

Proof: $\phi(x) = \left\{ 1, x, \frac{1}{\sqrt{2!}} x^2, \frac{1}{\sqrt{3!}} x^3 \cdots \right\}$.

$\langle \phi(x), \phi(z) \rangle = \sum_i \frac{x^i \cdot z^i}{i!} = e^{x^T z}$.

(Remember: We are talking about the inner product in $\phi(x)$, not x).