1

Bayes Decision Theory

Spring 2013

Note Title 9/30/2008

How to make decisions in the presence of uncertainty?

History: 2nd World War

Radar for detection aircraft.

Codebreaking. Decryption.

Observed Data $x \in \mathcal{X}$

State $y \in \mathcal{Y}$. likelihood function

$p(x|y)$ — conditional distribution

model how data is generated.

Example $y \in \{-1, 1\}$ Salmon / Sea Bass

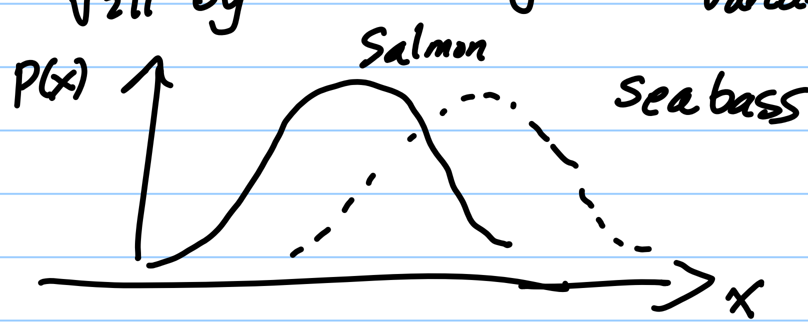
Airplane / Bird

$$p(x|y) = \frac{1}{\sqrt{2\pi}\,\sigma_y} e^{-\frac{1}{2}\frac{(x-\mu_y)^2}{2\sigma_y^2}}$$

mean μ_y

variance σ_y^2.

E.G. x is length of fish.



(2) How to decide Sea Bass or Salmon? Spring 2013

Airplane or Bird

Maximum Likelihood (ML)

$$\hat{y}_{ML} = \text{ARG}\max_{y} p(x|y) \qquad \left(p(x|\hat{y}_{ML}) \geq p(x|y) \ \forall y \right)$$

If $p(x|y=1) > p(x|y=-1)$ decide $y=1$
otherwise $y=-1$

Equivalently $\log \frac{p(x|y=1)}{p(x|y=-1)} > 0$ log-likelihood test.

Seems reasonable, but what if birds are more likely than airplanes?
Must take into account the prior probability $p(y=1), p(y=-1)$.

Bayes Rule $\quad p(y|x) = \frac{p(x|y)\,p(y)}{p(x)}$

prob of y conditioned on observation.

If $p(y=1|x) > p(y=-1|x)$ decide $y=1$
otherwise decide $y=-1$

Maximum a Posteriori (MAP) $\quad \hat{y}_{MAP} = \text{ARG}\max_{y} p(y|x)$

Sping 2013

(3) <u>Another ingredieid</u>

→ what does it cost if you make a mistake?

i.e. suppose you decide $y=1$, but really $y=-1$.

i.e. you may pay a big penalty if you decide it is a bird when it is a plane.

(Pascal's Wager: Bet on God)

Putting everything together.

likelihood function $p(x|y)$ $x \in X, y \in Y$

prior $p(y)$

decision rule $\alpha(x)$ $\alpha(x) \in Y$

loss function $L(\alpha(x), y)$ cost of making decision $\alpha(x)$ when true state is Y.

<u>E.G.</u> $L(\alpha(x), y) = 0$, if $\alpha(x) = y$

$L(\alpha(x), y) = 1$, if $\alpha(x) \neq y$

All wrong answers penalized the same.

Spring 2013

(4) Risk

The **risk** of the decision rule $\alpha(x)$ is the expected loss.

$$R(\alpha) = \sum_{x,y} L(\alpha(x), y)\, P(x,y)$$

(Note integrate $\int dx$ if x is continuous)

Bayes Decision Theory says "pick the decision rule $\hat{\alpha}$ which minimizes the risk".

$$\hat{\alpha} = \underset{\alpha \in A}{\mathrm{ARGMIN}}\, R(\alpha), \quad R(\hat{\alpha}) \geq R(\alpha) \quad \forall\, \alpha \in A.$$

A = set of all decision rules

$\hat{\alpha}$ is Bayes Decision
$R(\hat{\alpha})$ is Bayes Risk.

Spring 2013

(5) Bayes Risk

Bayes Risk is the best you can do if: (a) you know $p(x|y)p(y)$ & $L(\cdot,\cdot)$

(b) you can compute $\hat{\alpha} = \arg\min_{\alpha} R(\alpha)$.

(c) you can afford the losses (e.g. gambling, poker)

(d) you make the decision for a sequence of data $x_1 \ldots x_n$ with states $y_1 \ldots y_n$ where each (x_i, y_i) are independently identically distributed from $p(x,y)$

Bad – if you are playing a game against an intelligent opponent (Game Theory)

– if any of the assumptions (a), (b), (c), (d) are wrong.

Note: Cognitive Scientists have studied decision theory to see if it predicts the way humans make decisions. Results are debatable. But Prospect Theory (Kahneman, Tversky) suggests that humans do not.

Spring 2013

(6) Better understanding of Bayes Decision Theory. Re-express

$$R(\alpha) = \sum_x \sum_y L(\alpha(x), y) P(x, y)$$

$$= \sum_x P(x) \{ \sum_y L(\alpha(x), y) P(y|x) \},$$

Hence, for each x,

$$\hat{\alpha}(x) = \arg\min_{\alpha(x)} \sum_y L(\alpha(x), y) P(y|x)$$

Obtaining MAP & ML as special cases.

If $y \in \{-1, 1\}$ and the loss function penalizes all errors equally:

$$L(\alpha(x), y) = 1, \text{ if } \alpha(x) \neq y.$$
$$= 0, \text{ otherwise}$$

$y \in \{-1, 1\}$

Then $\hat{\alpha}(x) = \arg\max_{\alpha(x)} P(y = \alpha(x) | x)$ MAP estimate.

If also $p(y=1) = p(y=-1)$, then
$\hat{\alpha}(x) = \arg\max_{\alpha(x)} p(x | y = \alpha(x))$ ML estimate

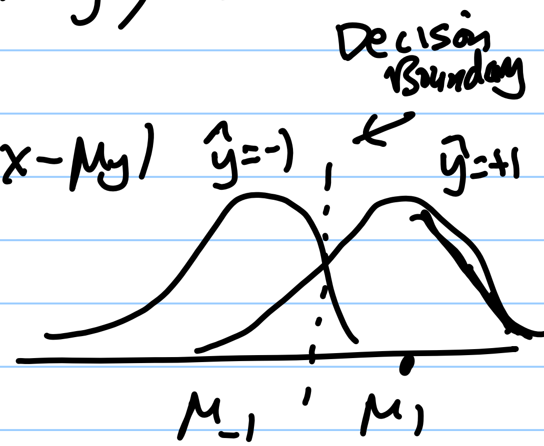
Spring 2013

(7) <u>Examples</u> $p(x|y) = \frac{1}{\sqrt{2\pi}\,\sigma_y} e^{-\frac{(x-\mu_y)^2}{2\sigma^2}}$.

$y \in \{-1, 1\}$ $p(y) = \frac{1}{2}$.

$L(\alpha(x), y) = 1$, if $\alpha(x) \neq y$, $= 0$ otherwise.

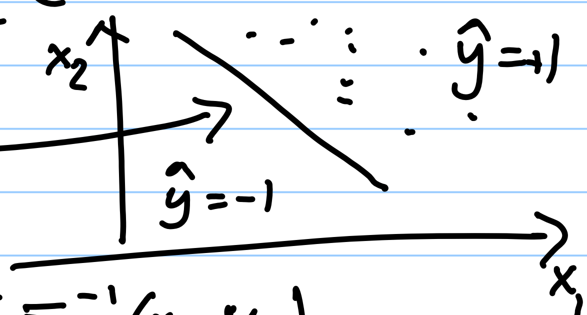
Bayes Rule $\alpha(x) = \arg\min_{y \in \{-1,1\}} |x - \mu_y|$



<u>Suppose</u> $\underline{x}$ is a vector in two dimensions

$$p(\underline{x}|y) = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}|\underline{x}-\underline{\mu}_y|^2}$$

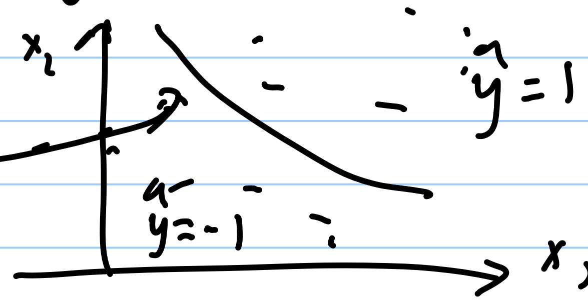
Seperating plane. $\leftrightarrow$ Decision Boundary



if $p(\underline{x}|y) = \frac{1}{2\pi|\underline{\underline{\Sigma}}_y|^{1/2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu}_y)^T \underline{\underline{\Sigma}}_y^{-1}(\underline{x}-\underline{\mu}_y)}$

Gaussians with unequal covariance

Decision Boundary



Spring 2013

More Details

$$P(\underline{x}|y) = \frac{1}{2\pi|\underline{\underline{\Sigma}}|^{1/2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu}_y)^T \underline{\underline{\Sigma}}^{-1}(\underline{x}-\underline{\mu}_y)}$$

ie same covariance $\underline{\underline{\Sigma}}$ for both classes $y=\pm 1$.

$$\log \frac{P(\underline{x}|y=1)}{P(\underline{x}|y=-1)} = \frac{1}{2}(\underline{x}-\underline{\mu}_{-1})^T \underline{\underline{\Sigma}}^{-1}(\underline{x}-\underline{\mu}_{-1}) - \frac{1}{2}(\underline{x}-\underline{\mu}_1)^T \underline{\underline{\Sigma}}^{-1}(\underline{x}-\underline{\mu}_1)$$

($2\pi|\Sigma|^{1/2}$ terms cancel)

Linear in $\underline{x}$ describes a plane.

$$= (\underline{\mu}_{-1}-\underline{\mu}_1)^T \underline{\underline{\Sigma}}^{-1}\underline{x} + \frac{1}{2}\underline{\mu}_{-1}^T \underline{\underline{\Sigma}}^{-1}\underline{\mu}_{-1} - \frac{1}{2}\underline{\mu}_1^T \underline{\underline{\Sigma}}^{-1}\underline{\mu}_1 .$$

Hence ML rule/estimator corresponds to a rule.

Classify $\underline{x}$ as $y=1$ if

$$(\underline{\mu}_{-1}-\underline{\mu}_1)^T \underline{\underline{\Sigma}}^{-1}\underline{x} + \frac{1}{2}\underline{\mu}_{-1}^T \underline{\underline{\Sigma}}^{-1}\underline{\mu}_{-1} - \frac{1}{2}\underline{\mu}_1^T \underline{\underline{\Sigma}}^{-1}\underline{\mu}_1 > 0$$

as $y=-1$ if $(\underline{\mu}_{-1}-\underline{\mu}_1)^T \underline{\underline{\Sigma}}^{-1}\underline{x} + \frac{1}{2}\underline{\mu}_{-1}^T \underline{\underline{\Sigma}}^{-1}\underline{\mu}_{-1} - \frac{1}{2}\underline{\mu}_1^T \underline{\underline{\Sigma}}^{-1}\underline{\mu}_1 < 0.$

If there is a prior $p(y)$

$$\log \frac{P(y=1|\underline{x})}{P(y=-1|\underline{x})} = \log \frac{P(\underline{x}|y=1)\,P(y=1)}{P(\underline{x}|y=-1)\,P(y=-1)}$$

$$= \log \frac{P(\underline{x}|y=1)}{P(\underline{x}|y=-1)} + \log \frac{P(y=1)}{P(y=-1)}$$

← Indep of $\underline{x}$

($P(y|\underline{x}) = \frac{p(\underline{x}|y)p(y)}{p(\underline{x})}$, $p(\underline{x})$ cancels in the ratio.)

Hence prior shifts the seperating plane to

$$(\underline{\mu}_{-1}-\underline{\mu}_1)^T \underline{\underline{\Sigma}}^{-1}\underline{x} + \frac{1}{2}\underline{\mu}_{-1}^T \underline{\underline{\Sigma}}^{-1}\underline{\mu}_{-1} - \frac{1}{2}\underline{\mu}_1^T \underline{\underline{\Sigma}}^{-1}\underline{\mu}_1 + \log\frac{P(y=1)}{P(y=-1)} .$$

With Loss Function

$$R(\alpha(\underline{x})=1) = L(1,1)P(y=1|\underline{x}) + L(1,-1)P(y=-1|\underline{x})$$

$$R(\alpha(\underline{x})=-1) = L(-1,1)P(y=1|\underline{x}) + L(-1,-1)P(y=-1|\underline{x})$$

Decision boundary occurs where $R(\alpha(\underline{x})=1) = R(\alpha(\underline{x})=-1)$.

ie, when $\{L(1,1)-L(-1,1)\}P(y=1|\underline{x}) = \{L(-1,1)-L(1,-1)\}P(y=-1|\underline{x})$

ie when $\log\frac{P(y=1|\underline{x})}{P(y=-1|\underline{x})} = \log\frac{\{L(-1,1)-L(1,-1)\}}{\{L(1,1)-L(-1,1)\}}$

↗ additional shift in position of seperating plane.

 Spring 2013

Bayes Decision theory also applies when y is not a binary variable – e.g. y can take M values or y continuous valued.

In this course, usually $y \in \{-1, 1\}$ classification

or $y \in \{1, 2 \dots M\}$ multi-class classification

or $y \in \{-\infty, \infty\}$ regression.

But there is a major problem with Bayes Decision Theory, apart from the limitations discussed earlier.

Problem: We almost never know the probability distributions $p(y|\underline{x})$ and $p(\underline{x})$.

Instead we have data $\{(\underline{x}_i, y_i) : i = 1 \text{ to } N\}$

E.G. Bank has records of the incomes and savings of the customers – $\underline{x}_i$. and whether they defaulted on their loans – y_i.

Similarly, you can make a dataset of fish, recording their length and brightness $\underline{x}_i$ and whether they are salmon or sea bass y_i

(10) 2013

Two Strategies

(1) Probability Approach.

Use the data $\{(\underline{x}_i, y_i) : i = 1 \text{ to } N\}$ to learn probability distribution $p(\underline{x}|y)$ and $p(y)$. Then apply Bayes Decision Theory.

E.G. $p(y=1) = \frac{\sum_{i=1}^{N} I(y_i=1)}{N}$ Indicator function: $I(y=1) = 1$, if $y=1$

$p(y=-1) = \frac{\sum_{i=1}^{N} I(y_i=-1)}{N}$ $I(y=1) = 0$, otherwise

Gaussian assumption $P(\underline{x}|y=1) = N(\underline{\mu}_1, \underline{\Sigma}_1)$ $N(\text{mean}, \text{covariance})$ normal

$P(\underline{x}|y=-1) = N(\underline{\mu}_{-1}, \underline{\Sigma}_{-1})$

$$\underline{\mu}_1 = \frac{\sum_{i=1}^{N} I(y_i=1)\underline{x}_i}{\sum_{i=1}^{N} I(y_i=1)}, \quad \underline{\mu}_{-1} = \frac{\sum_{i=1}^{N} I(y_i=-1)\underline{x}_i}{\sum_{i=1}^{N} I(y_i=-1)}$$

$$\underline{\Sigma}_1 = \frac{1}{\sum_{i=1}^{N} I(y_i=1)} \sum_{i=1}^{N} I(y_i=1)(\underline{x}_i-\underline{\mu}_1)(\underline{x}_i-\underline{\mu}_1)^T$$

$$\underline{\Sigma}_2 = \frac{1}{\sum_{i=1}^{N} I(y_i=-1)} \sum_{i=1}^{N} I(y_i=-1)(\underline{x}_i-\underline{\mu}_{-1})(\underline{x}_i-\underline{\mu}_{-1})^T$$

i.e. estimate the mean and covariances for classes $y=1$ and $y=-1$ using only the data assigned to that class (e.g. assign $\underline{x}_i$ to class $y=1$, if $y_i=1$).

Note: This strategy requires learning parametric and non-parametric probability distribution – we will discuss methods for doing this in later lectures.

(11) Strategy (2) Spring 2013

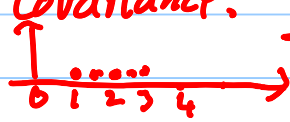
Discriminative : Attempt to learn the decision rule $\alpha(x)$ directly from the data $\{(\underline{x}_i, y_i) : i = 1 \text{ to } N\}$.

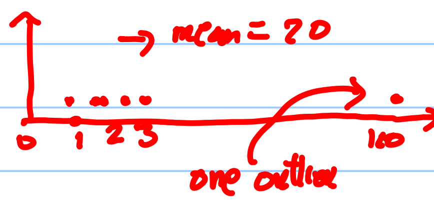
i.e. select $\hat{\alpha}(.) = \arg\min_{\alpha()} R_{emp}(\alpha : N)$
$= \arg\min_{\alpha(.)} \sum_{i=1}^{N} L(\alpha(\underline{x}_i), y_i)$

Justification – why bother learning the probabilities if we really only care about the decision rule?

E.G. Suppose we use Gaussians to model $P(\underline{x}|y)$.
Gaussians are non-robust – outliers in the data, i.e. atypical values of $\underline{x}_i$, can make big changes to the estimation of the mean and covariance.

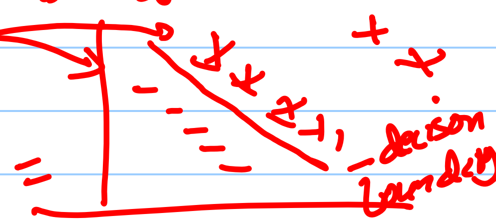
→ Data.





So if we to learn $P(x|y)$ our estimates of the mean and covariance, hence of the decision boundary, can be corrupted by outliers far away from the boundary.

But if instead, we just search for a linear plane that seperates the data from $y=1$ and $y=-1$ then we only need to pay attention to the data near the boundary



This is particularly helpful if we have little data, as is usually the case.

 Spring 2013

Key Issue of Machine Learning:

How to generalize — learn a classification rule on training data $\{(\underline{x}_i, y_i) : i = 1 \text{ to } N\}$ which also works on new data that you haven't seen.
i.e. predict on new data.

How to formalize this?

Assume that the training data $\{(\underline{x}_i, y_i) : i = 1 \text{ to } N\}$ are independently identically distributed (i.i.d) from an unknown distribution $P(\underline{x}, y)$.

Want the decision rule $\alpha(.)$ trained on $\{(\underline{x}_i, y_i) : i = 1 \text{ to } N\}$ to work on other i.i.d samples $\{(\underline{x}_j, y_j) : j = N+1 \text{ to } N+M\}$, from $P(\underline{x}, y)$

i.e. if empirical risk $R_{emp}(\alpha : N)$ is small then want the risk $R(\alpha)$ to be small.

Now $\alpha \in A$, where A is a set of decision rules [e.g. could include ML, MAP separating planes, nearest neighbor, decision trees].

So if we can make sure that

$$|R_{emp}(\alpha, N) - R(\alpha)| \text{ is small for all } \alpha \in A$$

then we can select a rule $\tilde{\alpha} = \arg\min_{\alpha} R(\alpha, N)$ and be confident that

$$R(\tilde{\alpha}) \text{ is close to } \min_{\alpha} R(\alpha)$$

ie that rule $\tilde{\alpha}$ generalizes.

(13) # Memorization vs. Generalization Spring 2013

$$R_{emp}(\alpha) = \frac{1}{N}\sum_{i=1}^{N} L(\alpha(x_i), y_i)$$

suppose $L(\alpha(x_i), y_i) \in \{1, 0\}$

By law of large numbers $R_{emp}(\alpha) \xrightarrow[N\to\infty]{} R(\alpha) = \sum_{x,y} p(x,y) L(\alpha(x), y)$

but how fast?

Fix α: By standard theorems (Chernoff, Sanov, Cremers.)

$$\Pr\{ |R_{emp}(\alpha) - R(\alpha)| > \epsilon \} < e^{-N\epsilon}$$

require $e^{-N\epsilon} < \delta \iff N > -\frac{1}{\epsilon}\log\delta$ {any ϵ} {$-\log\delta > 0$ if $0 \leq \delta \leq 1$}

So, if $N > -\frac{1}{\epsilon}\log\delta$, then with prob $> 1-\delta$

$|R_{emp}(\alpha) - R(\alpha)| < \epsilon$. Almost sure that we can

Probably Approximately Correct (PAC) estimate Bayes risk from N samples.

But, we have to consider many different rules α.

For simplicity, suppose we consider a finite no. of rules $\{\alpha^\nu : \nu = 1 \text{ to } H\}$

We want $|R_{emp}(\alpha^\nu) - R(\alpha^\nu)| < \epsilon$ to be small for all ν, with high probability.

Boole's Inequality: $\Pr(A^1 \text{ or } \ldots \text{ or } A^H) \leq \sum_{\nu=1}^{H} \Pr(A^\nu)$

Let $\Pr(A^\nu)$ be prob that $|R_{emp}(\alpha^\nu) - R(\alpha^\nu)| > \epsilon$

$\Pr\{$At least one rule A^ν has error greater than $\epsilon\}$

$< He^{-N\epsilon}$ Now want $He^{-N\epsilon} < \delta$

$\iff N > \frac{1}{\epsilon}\{\log H - \log\delta\}$

So if $N > \frac{1}{\epsilon}\{\log H - \log\delta\}$, then with prob $> 1-\delta$

$|R_{emp}(\alpha^\nu) - R(\alpha^\nu)| < \epsilon$ for all $\nu = 1$ to H.

Hence number of examples needed grows rapidly with H (size of hypothesis space), accuracy required ϵ, certainty δ.

(14) Spring 2013

Memorization:

Decision Rule: $\hat{\alpha} = \underset{\alpha}{\text{ARGMIN}}\ R_{emp}(\alpha)$

$R_{emp}(\hat{\alpha})$ small, but $R(\alpha)$ big.

i.e. bad for predicting new data.

Generalization:

Want a decision rule $\hat{\alpha}$ so that

$R_{emp}(\hat{\alpha})$ is small, but $R(\hat{\alpha})$ is small.

In practice — cross-validation.

training set $\{(x_i, y_i) : i = 1 \text{ to } N\}$

to learn the rule $\bar{\alpha}$

test set $\{(x_j, y_j) : j = 1 \text{ to } M\}$

to test the rule $\bar{\alpha}$.

Choose $\hat{\alpha}$ so that $R_{emp}(\bar{\alpha})$ is small on both the training set and test set.

How, restrict the possibilities of $\hat{\alpha}$.

(15) Spring 2013

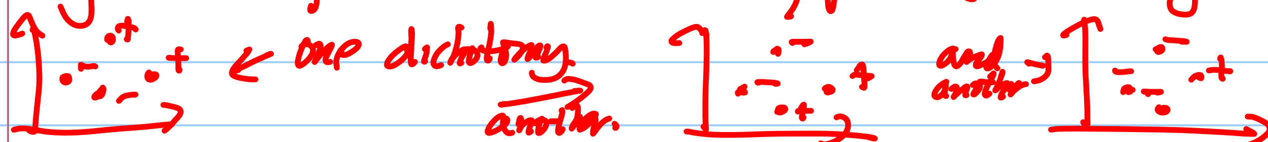
What happens if we have an infinite set of rules? — e.g. the set all seperating planes $ax+by+c=0$

The Vapnik-Chervonenkis VC dimension give a finite measure of the capacity of a hypothesis class A.

Introduce the concept of shattering.

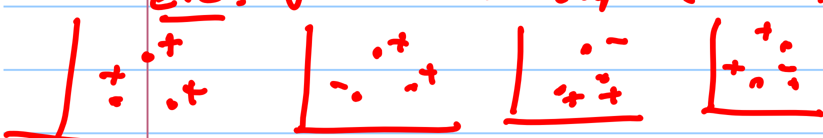
Suppose we have n data examples (features/attributes) $\{\underline{x}_i : i=1..n\}$ in d-dim space. With general position assumption (data doesn't lie on a lower-dimensional subspace).

There are 2^n possible dichotomies of the data — separating the examples into two classes, positive and negative

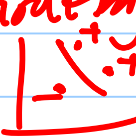


A set A of classifiers, shatters n examples in d-dim space if, for all dichotomies of the data, we can find a classifier in A which classifies the data correctly.

E.G. If we have 3 datapoints in 2D, there are $2^3=8$ dichotomies.



and some examples with + replaced by −

For each dichotomy, we can find a separating plane which classifies the data perfectly → eg 

Hence, we know that we can classify the data perfectly before we even look at it.

(16) The VC-dimension of a hypothesis class A is the maximum number of points that can be shattered. Note: this depends on the dimension of the space.

For seperating hyperplanes, the VC dimension $= d+1$. i.e. VC = 3 for
$\quad\quad\uparrow$ dim of space. planes in 2D space.

This concept enables us to prove theorems for hypothesis spaces with finite VC dimension, but infinite number of classifiers (e.g. planes)

For example,

with prob $> 1-\delta$

$$R(\alpha) \leq R_{emp}(\alpha : N) + \sqrt{\frac{h(\log 2N/h) - \log \delta/4}{N}} \quad \text{for all } \alpha \in A$$

PAC Theorem where h is the VC dimension of A
N is the total amount of data.

Moral: In order to generalize, you have to restrict the complexity (i.e. the VC dimension) of the set of classifiers you use by taking into account the amount of data