(1.) # Bayes Decision Theory Spring 2014

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

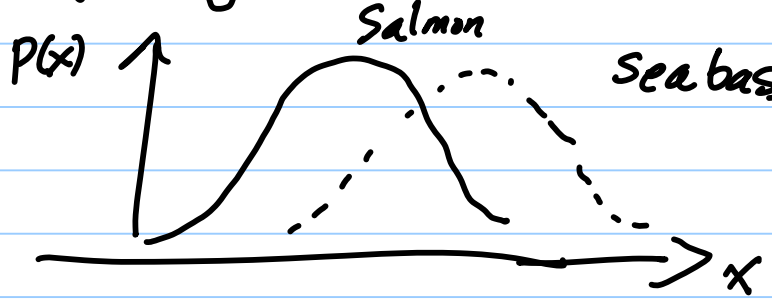
$p(x|y)$ — conditional distribution
model how data is generated.

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

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

mean μ_y
variance σ_y^2.

E.G. x is length of fish.



(2) How to decide Sea Bass or Salmon? Airplane or Bird

Maximum Likelihood (ML)

$$\hat{y}_{ML} = \arg\max_y p(x|y)$$

$$\left(p(x|\hat{y}_{ML}) \geq p(x|y) \quad \forall y \right)$$

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 $p(y|x) = \frac{p(x|y)p(y)}{p(x)}$

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) $\hat{y}_{MAP} = \arg\max_y p(y|x)$

(3) Another Ingredient

→ 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 $p(x|y)$ $\quad x \in X, y \in Y$

prior $p(y)$

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

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

e.g. $L(\alpha(x), y) = 0$, if $\alpha(x) = y$

$L(\alpha(x), y) = 1$, if $\alpha(x) \neq y$, all errors penalized the same.

or $L(\alpha(x), y) = 0$, if $\alpha(x) = y$

$L(\alpha(x)=1, y=-1) = 10$

$L(\alpha(x)=-1, y=1) = 10{,}000{,}000{,}000{,}000$ PASCAL'S CASE.

$y=1$, God exists, $y=-1$, God does not exist.

(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 $\alpha(\cdot)$ which minimizes the risk".

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

A = set of all decision rules

$\hat{\alpha}$ is Bayes Decision

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

(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 $\hat{d} = \arg\min_d R(d)$

(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. (use Game Theory instead)

Bad – if any of (a), (b), (c), (d) are wrong

Note: Cognitive Scientists study whether people use decision theory. Kahneman & Tversky argue that people do not – Prospect Theory. Debatable.

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

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

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

Hence, for each x,

$$\hat{\alpha}(x) = \underset{\alpha(x)}{\text{ARG MIN}} \sum_y L(\alpha(x), y) P(y|x)$$

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 $\hat{\alpha}(x) = \underset{\alpha(x)}{\text{ARG MAX}} P(y = \alpha(x) | x)$ MAP estimate.

If also $P(y=1) = P(y=-1)$, then

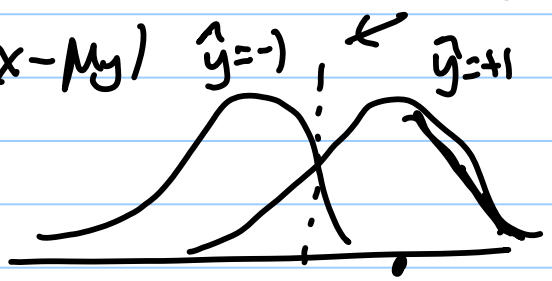
$$\hat{\alpha}(x) = \underset{\alpha(x)}{\text{ARG MAX}} p(x | y = \alpha(x))$$ ML estimate

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

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

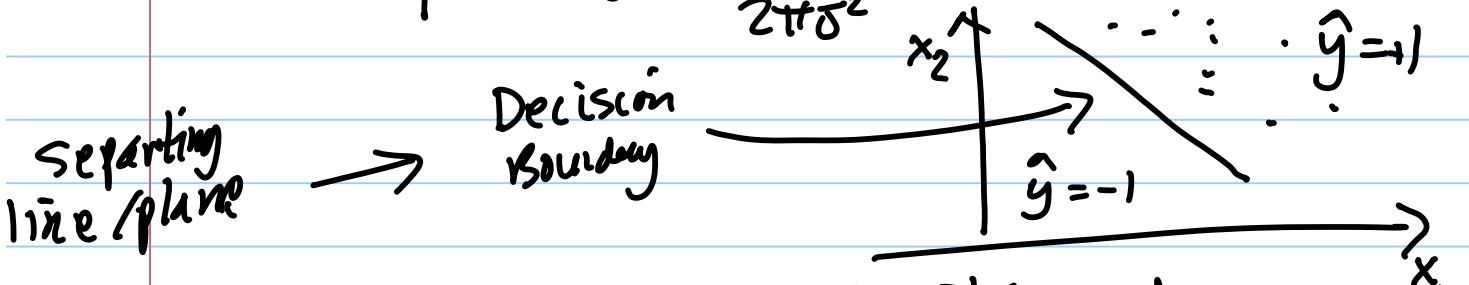
$L(\alpha(x), y) = 1$, if $\alpha(x) \neq y$, $= 0$ otherwise.

Bayes Rule $\alpha(x) = \underset{y \in \{-1,1\}}{\text{ARG MIN}} |x - \mu_y|$



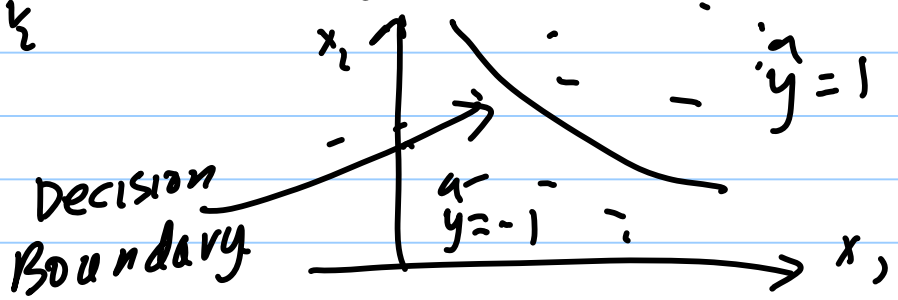
Suppose $\underline{x}$ is a vector in two dimensions

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



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

Gaussians with unequal covariance.



More Details

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

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

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

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

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

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

Hence ML rule/estimator corresponds to a rule.

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

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

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

If there is a prior $p(y)$

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

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

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

$\leftarrow$ Indep of $\underline{x}$

Hence prior shifts the separating plane to

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

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 $R(\alpha(x)=1) = R(\alpha(x)=-1)$.

is, 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 separating plane.

(9)

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

(i) $y \in \{-1, 1\}$ binary classification.

(ii) $y \in \{1, 2 \ldots M\}$ multi-class classification.

(iii) $y \in (-\infty, \infty)$ regression.

Note: machine learning also addresses cases where $\underline{y} = (y_1, y_2 \ldots, y_N)$ is a vector

$y_i \in \{\pm 1\}$

$y_i \in \{1, 2 \ldots M\}$

$y \in (-\infty, \infty)$

but this is beyond the scope of this course.

(16) Problem (a). Bayes Decision Theory

We usually do not know the distributions $p(y|\underline{x})$ and $p(\underline{x})$

Instead we know data $\mathcal{X}_N = \{(\underline{x}_i, y_i) : i = 1, \ldots, N\}$

E.G. We have bank records of income and savings of N customers, and know if they defaulted or not. $\{(\underline{x}_i, y_i) : i = 1, \ldots, N\}$, $\underline{x}$: income, savings; y: defaulted or not

Key Assumption

In any Machine Learning problem we assume that the data we observe is generated by an (unknown) probability distribution

The data examples

$(\underline{x}_1, y_1), (\underline{x}_2, y_2), \ldots, (\underline{x}_N, y_N)$

are independent, identically distributed (iid) samples from $p(\underline{x}, y)$.

We want to obtain a decision rule $y = \alpha(\underline{x})$ which is good (i.e. small risk) for all possible samples from $p(\underline{x}, y)$ (generalization). A decision rule which has low risk for the data examples (memorization) is not good enough.

(11)

This suggests two strategies.

Strategy (1): The Probabilistic Approach.

Use the data $\{(\underline{x}_i, y_i) : i = 1 \text{ to } N\}$ to learn probability distributions $P(\underline{x}|y)$ 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

E.G. Gaussian assumption $P(\underline{x}|y=1) = \mathcal{N}(\underline{\mu}_1, \underline{\Sigma}_1)$ $\mathcal{N}(\cdot)$ = normal = common

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

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

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

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

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 distributions — we will discuss methods for doing this in later lectures.

(12) Strategy (2): Decision Rule

Learn the decision rule $y = \alpha(x)$ directly.

Define the empirical risk $R_{emp}(\alpha, \chi_N) = \frac{1}{N} \sum_{i=1}^{N} L(\alpha(x_i), y_i)$.

This depends on the dataset: $\chi_N = \{(x_i, y_i) : i = 1 \text{ to } N\}$

For example (but not always - see later lectures).

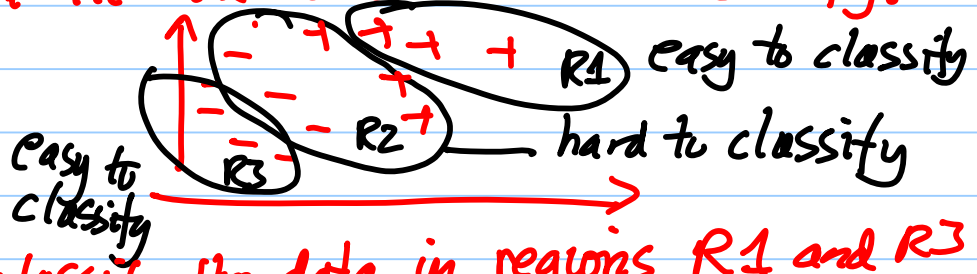
select $\hat{\alpha}(\,) = \arg\min_{\alpha} R_{emp}(\alpha; \chi_N)$

Motivation: (i) why bother to learn the probability distributions if your final goal is to obtain the decision rule?

(ii) you may make mistakes when you learn the probability distributions - because you will have to make assumptions about them (see next lectures) which may be incorrect. (see example on next page)

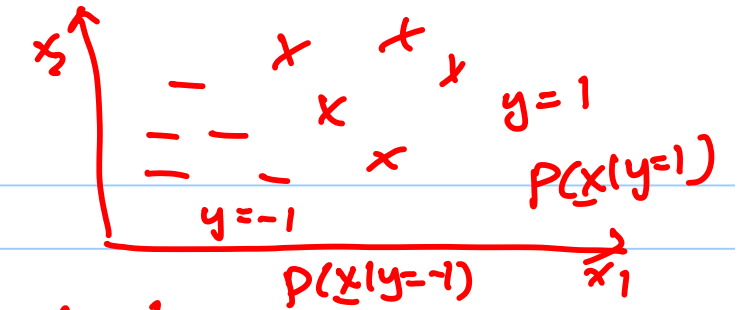
(iii) you should concentrate your effort by dealing with the data which is hard to classify.

Example:



It is easy to classify the data in regions R1 and R3 (because these regions contain only +ve and -ve examples respectively) So we should concentrate our effort in R2 (+ve and -ve examples) But the probability strategy would pay equal attention to R1, R2, R3

(13) Danger of Using Probabilities

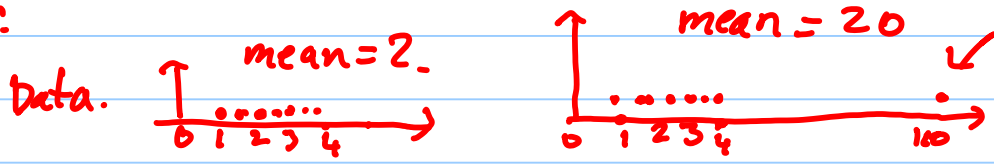


Suppose we assume the distributions $P(\underline{x}|y)$ are Gaussians

But Gaussians are <u>non-robust</u>.

Outliers in the data — unlikely values of $\underline{x}$. — can make big changes to the estimates of the means and covariances of the Gaussians.

<u>Example:</u>

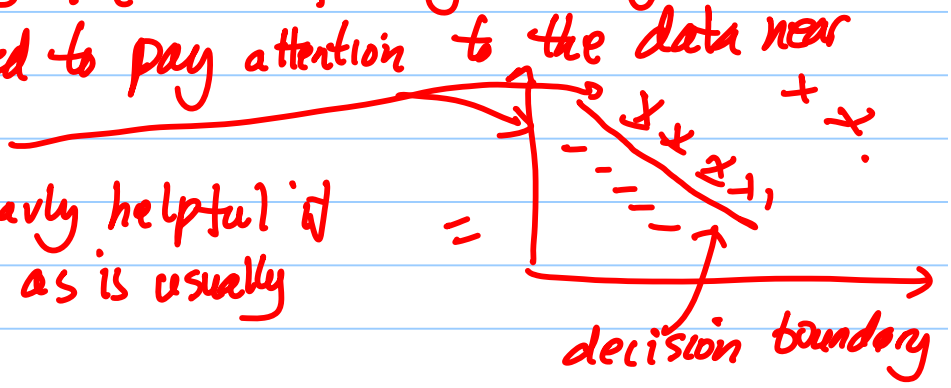


<u>Note</u>: we have to assume a model for $p(\underline{x}|y)$ when we learn from data - if model is wrong, then mistakes happen.

So if we try to learn $P(\underline{x}|y)$ 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.



(14) Fundamental Problem of Machine Learning

We want to find $\tilde{\alpha}(.)$ to minimize Bayes Risk $R(\alpha)$
– this is generalization.

But we only know the empirical risk $R_{emp}(\alpha; \mathcal{X}_N)$ of the dataset $\mathcal{X}_N = \{(\underline{x}_i, y_i) : i = 1, \ldots, N\}$.

Finding $\tilde{\alpha}(.)$ to minimize $R_{emp}(\alpha; \mathcal{X}_N)$ may only memorize the dataset, which is not what we want.

Fundamental Assumption: the dataset $\mathcal{X}_N$ consists of i.i.d. samples from $P(\underline{x}, y)$.

Insight. As the dataset gets bigger $N \to \infty$, the empirical risk converges (in probability) to the Bayes risk. i.e. $R_{emp}(\alpha; \mathcal{X}_N) \to R(\alpha)$

Let $\mathcal{A}$ be the set of all decision rules, (e.g. ML, MAP, seperating planes, nearest neighbor, decision trees).

Now suppose N is large enough so that $|R_{emp}(\alpha; \mathcal{X}_N) - R(\alpha)|$ is small, for all $\alpha \in \mathcal{A}$

then we can select a rule $\tilde{\alpha} = \underset{\alpha \in \mathcal{A}}{\arg\min}\, R(\alpha, N)$

and be confident that

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

i.e. that the rule $\tilde{\alpha}$ works well on all data from $P(\underline{x}, y)$, that it generalizes.

How big should N be? It depends on the set $\mathcal{A}$ of decision rules. This is Advanced Material.

(15) Memorization vs. Generalization Advanced Material

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

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, Cramer.)

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

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

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

$|R_{emp}(\alpha) - R(\alpha)| < \epsilon$ → Almost certain we can estimate the risk of α from $N > \frac{-1}{\epsilon}\log\delta$ examples

Probably Approximately Correct (PAC)

But, we must consider many different rules $\alpha \in A$

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

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

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

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

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

$< H e^{-N\epsilon}$

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

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

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

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

Hence number of examples needed grows rapidly with H size of hypothesis space

accuracy required ϵ, certainty δ.

(16) 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 $\bar{\alpha}$ so that $R_{emp}(\bar{\alpha})$ is small, but $R(\bar{\alpha})$ 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 $\bar{\alpha}$ so that $R_{emp}(\bar{\alpha})$ is small on both the training set and test set.

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

Advanced Material

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

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

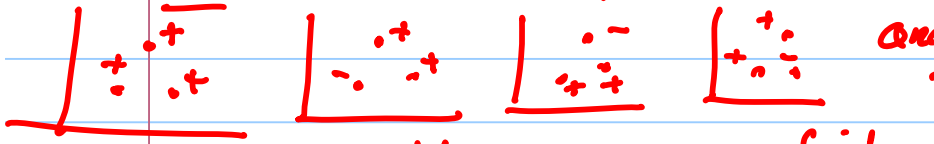
Introduce the concept of shattering.

Suppose we have n data examples (features/attributes) $\{x_i : i=1,\dots,n\}$ 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.



For each dichtomy, 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.

Advanced Material

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 $= d+1$ (where d = dim of space). i.e. VC = 3 for 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