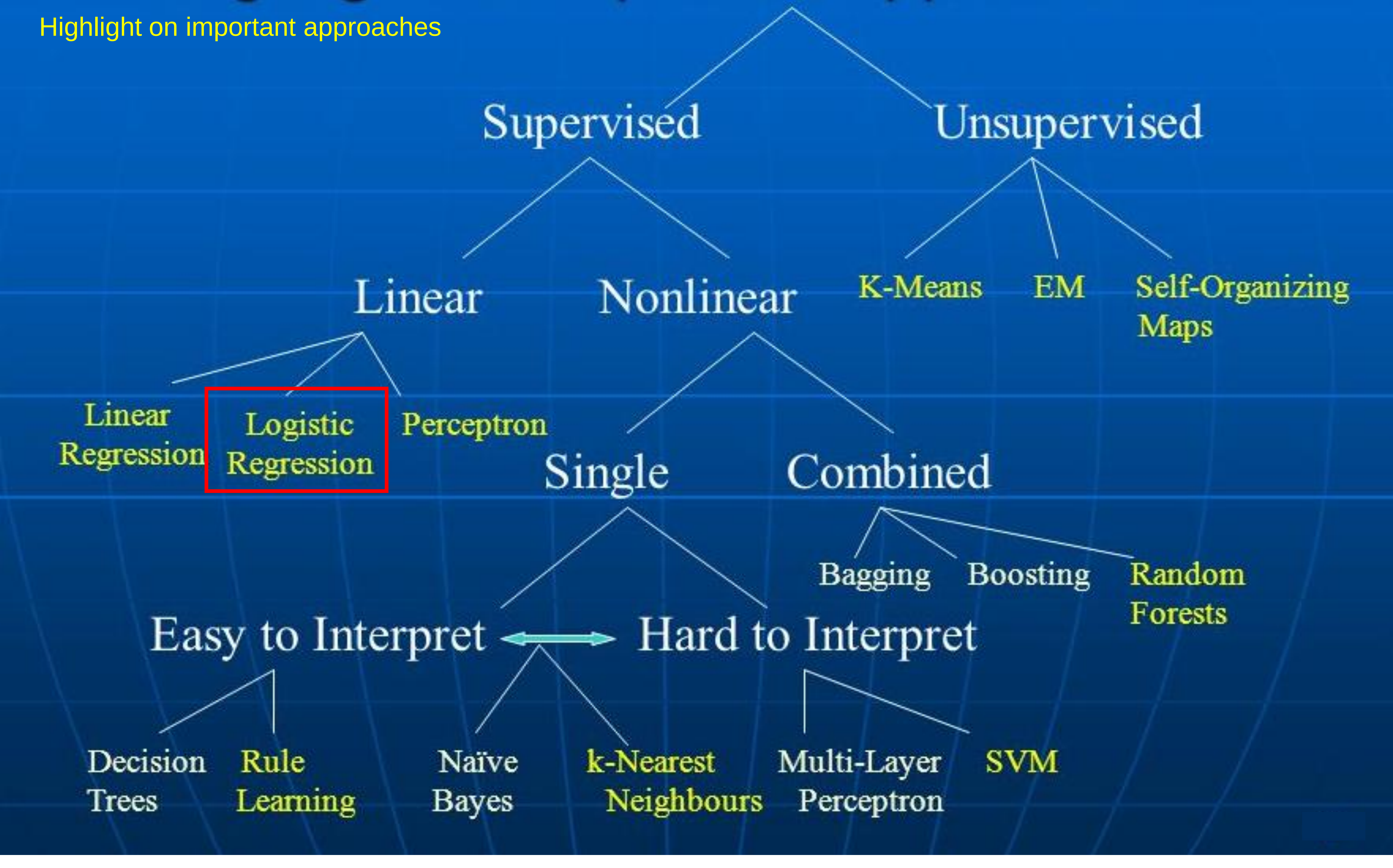
로지스틱회귀분석(Logistic Regression)

특허법인 가산
채승원 변리사(csw@kspat.com)

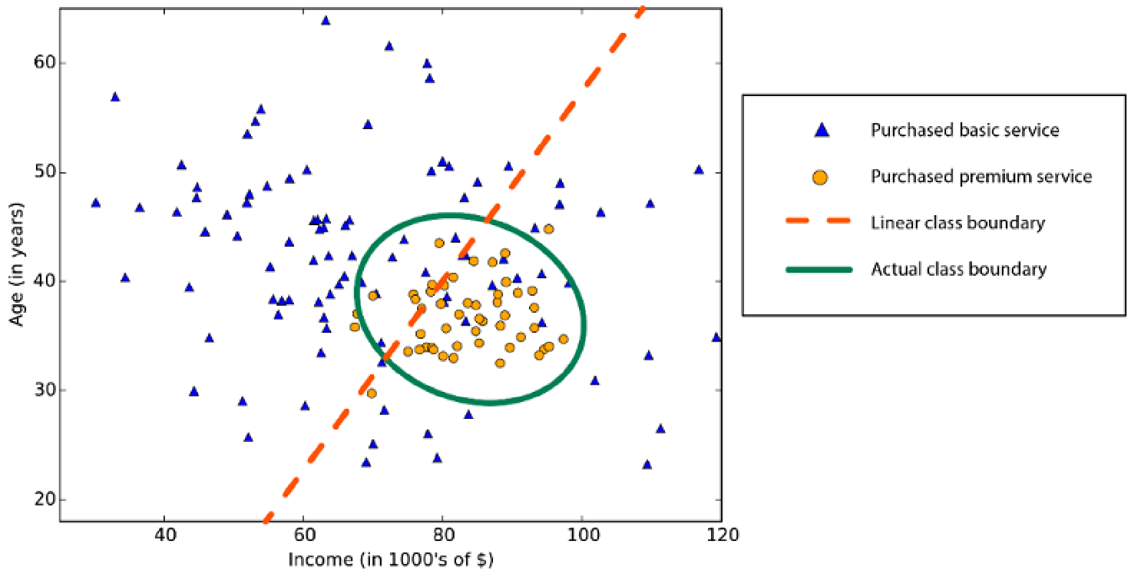
Machine Learning Taxonomy

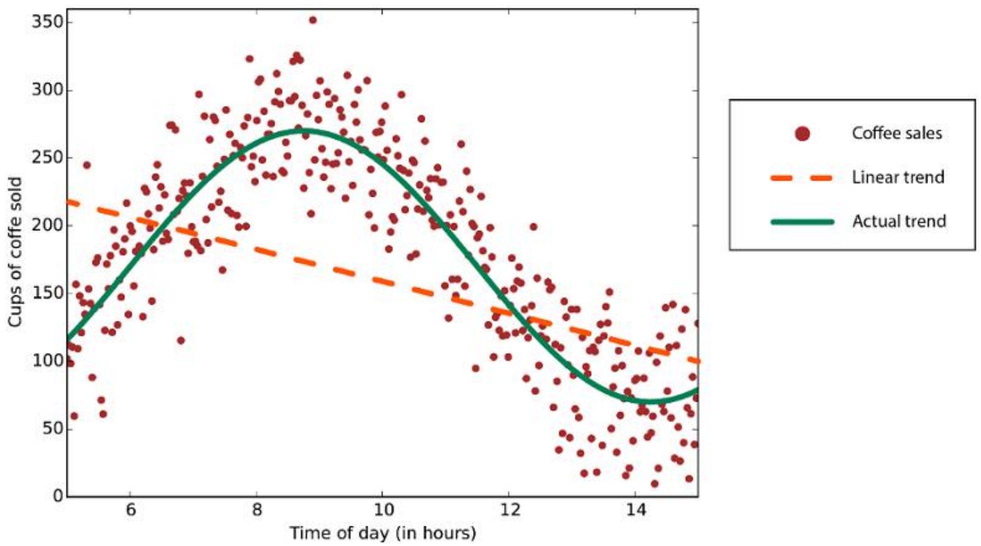
Highlight on important approaches



Linearity
Lots of machine learning algorithms make use of linearity. Linear classification algorithms assume that classes can be separated by a straight line (or its higher-dimensional analog). These include logistic regression and support vector machines. Linear regression algorithms assume that data trends follow a straight line. These assumptions aren't bad for some problems, but on others they bring accuracy down.

Non-linear class boundaries





The Logistic regression classifier

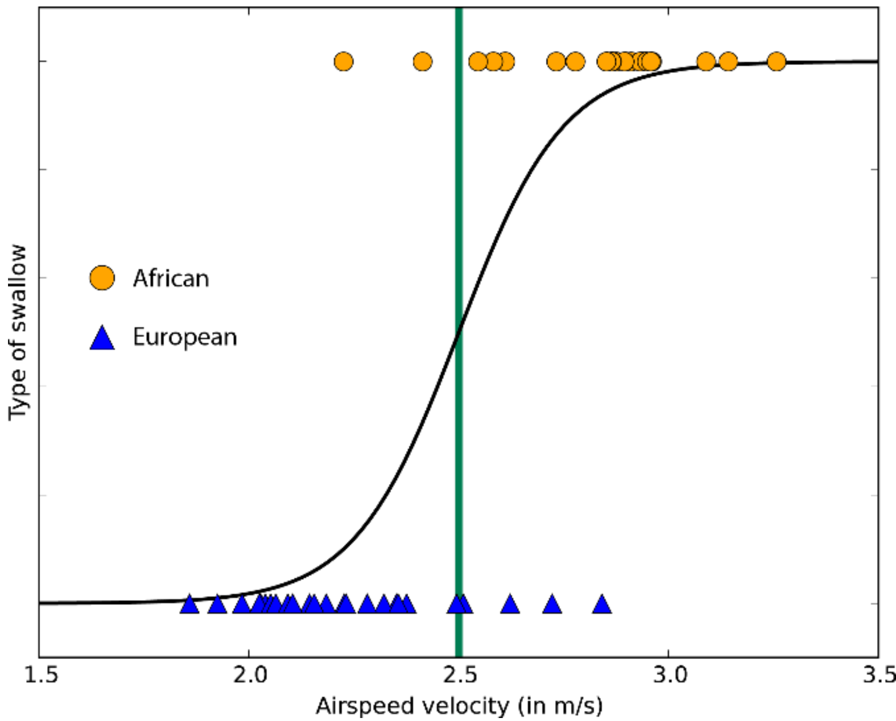
Regression은 연봉, 지원 수, 수학 점수 등 continuous한 값을 추정하기 위한 것. BUT! binary(TRUE/FALSE) value를 추정하기 위하여는 어떤 regression을 이용해야 하는가?

→ LOGISTIC REGRESSION

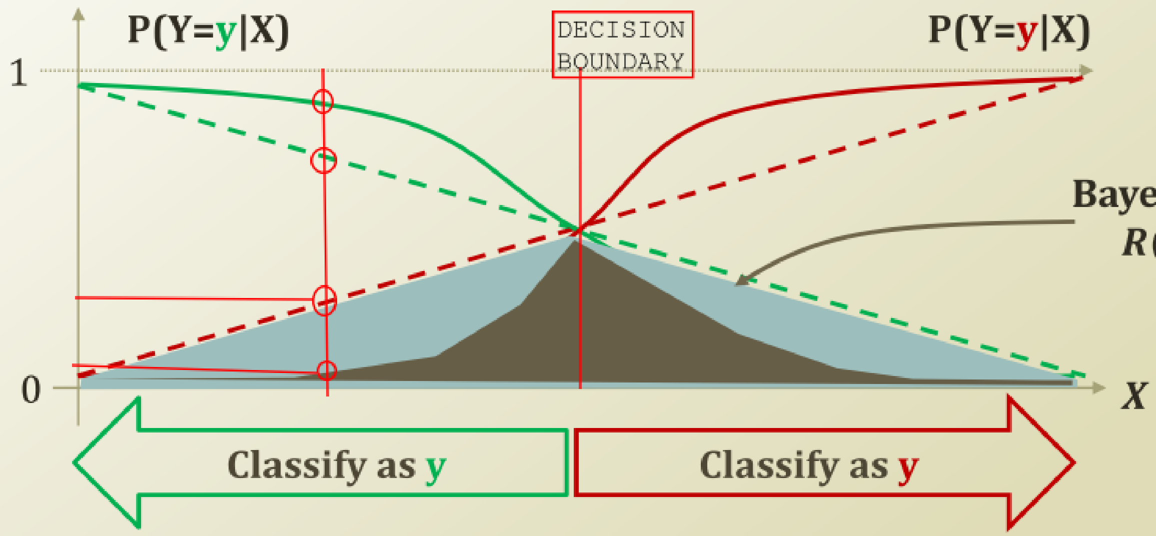
→ Naïve assumption 없이 classification 할 수 있는 방법 중 하나

→ 로지스틱 회귀는 선형 회귀 분석과는 다르게 종속 변수가 범주형 데이터를 대상으로 하며 입력 데이터가 주어졌을 때 해당 데이터의 결과가 특정 분류로 나눠지기 때문에 일종의 분류 (classification) 기법

→ multinomial에 대하여도 적용 가능

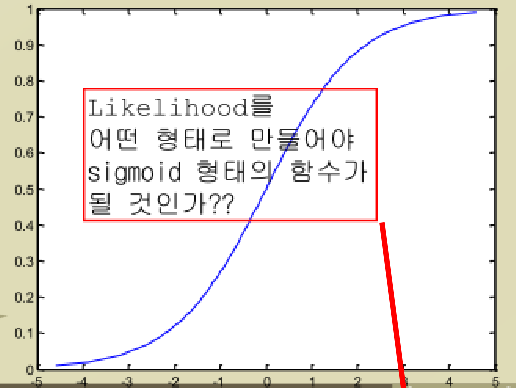


Optimal Classification and Bayes Risk



- Linear function vs. Non-linear function of P(Y|X)
  - Which is better?
- Problems of linear function
  - Range
  - Risk optimization
- Which function to use?
  - Need S-curve!

S-curve
a.k.a. Sigmoid function



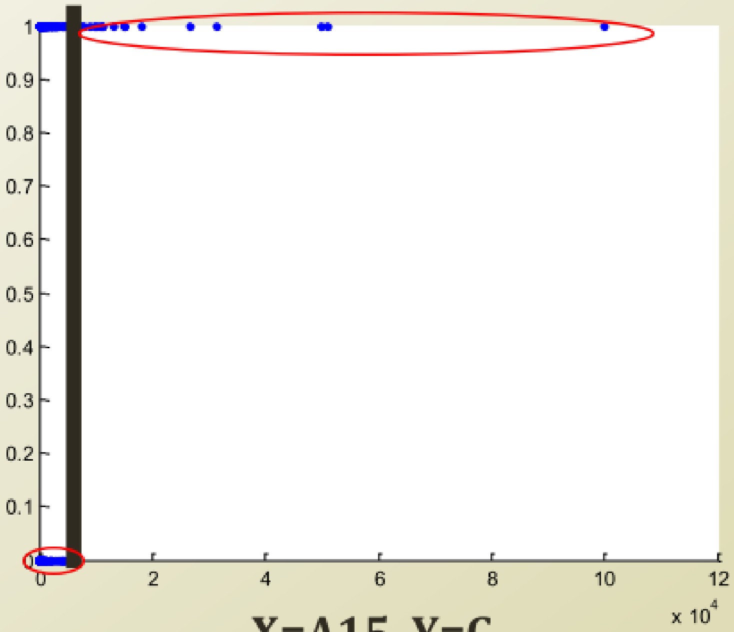
- 로지스틱 모형: $g(x) = \dfrac{e^x}{1+e^x}$
- 검벨 모형: $g(x) = e^{-e^x}$

이 중에 계산상의 편리성으로 인하여 로지스틱 모형이 널리 사용된다.

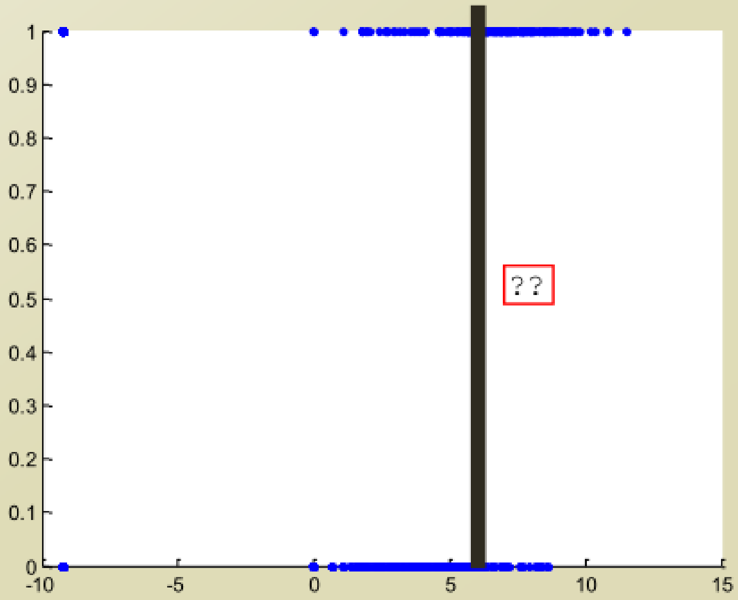
Classification with One Variable

true(1) / false(0)

- Let's predict the class, C, with an attribute, A15
  - Imagine that the Y axis shows P(Y|X)
  - There is a decision boundary
    - You can see it intuitively
- Then, How to find the boundary?



X=A15, Y=C



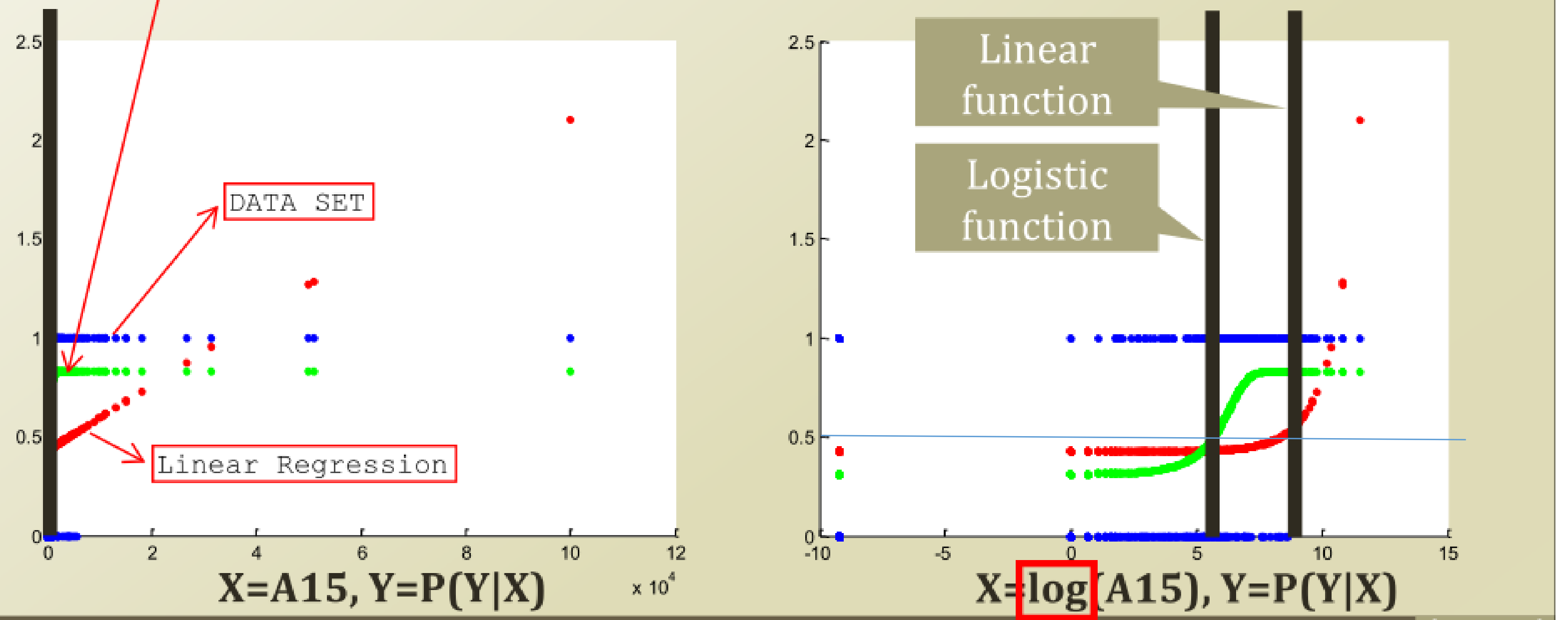
X=log(A15), Y=C

The Logistic regression classifier

Linear Function vs. Non-Linear Function

- Problem of fitting to the linear function
  - Violate the probability axiom
  - Slow response to the examples
- Better to fit to the logistic function
  - Keep the probability axiom
  - Quick response around the decision boundary
- Which function to use?
  - Logistic function – a special case of sigmoid function

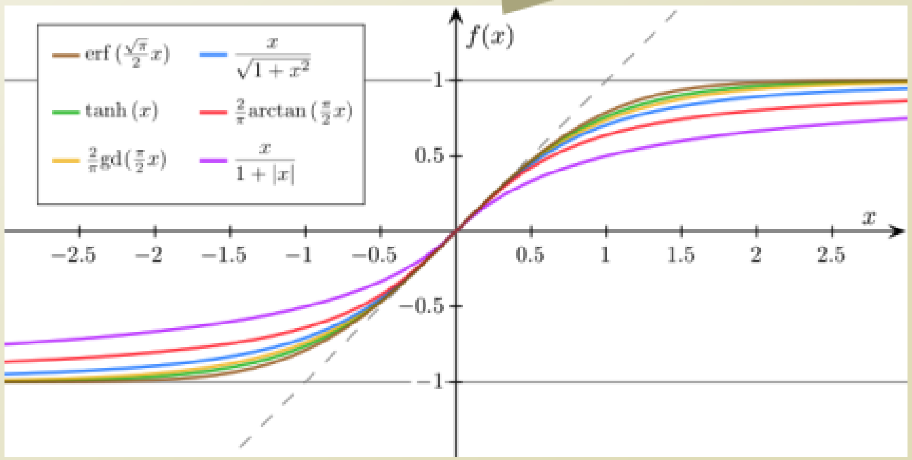
Blue = (X, Y_{true})
Red = $(X, P_{lin}(Y|X))$
Green = $(X, P_{log}(Y|X))$



Logistic Function의 Decision boundary가 Linear function의 그것 보다 정확

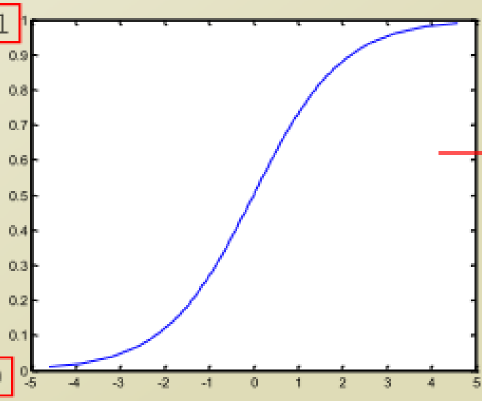
Logistic function

Many types of sigmoid functions



- Sigmoid function is
  - Bounded
  - Differentiable
  - Real function
  - Defined for all real inputs
  - With positive derivative
- Logistic function is
  - $\boxed{f(x) = \frac{1}{1+e^{-x}}}$
  - In relation to the population growth
  - Why is this good?
    - Sigmoid function
    - Particularly, easy to calculate the derivative…

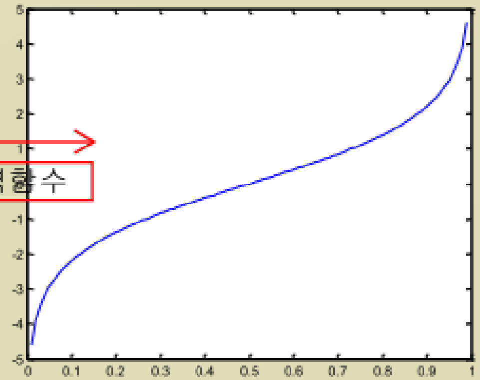
1



0

Logistic Function

역함수

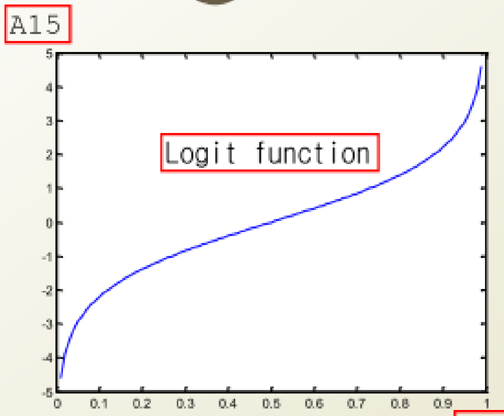


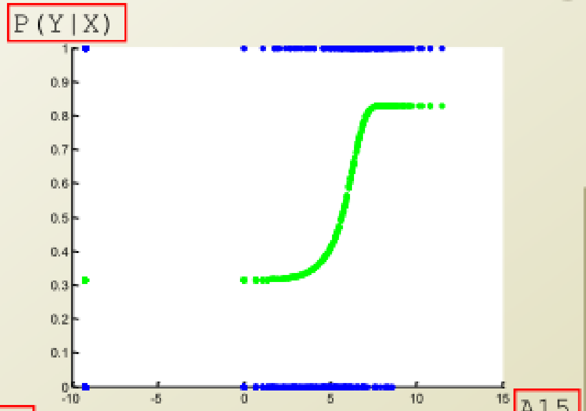
Logit Function

$$f(x) = \log\left(\frac{x}{1-x}\right)$$

미분하여 0이 되는 값을 구하기 편함.

Logistic Function Fitting





Linear Regression:

$\hat{f} = X\theta \quad \theta = (X^T X)^{-1} X^T Y$

Very similar to the linear regression.
Turning to the multivariate case

$$f(x) = \log\left(\frac{x}{1-x}\right) \rightarrow \boxed{x} = \log\left(\frac{p}{1-p}\right) \rightarrow \boxed{ax+b} = \log\left(\frac{p}{1-p}\right) \rightarrow \boxed{X\theta} = \log\left(\frac{p}{1-p}\right)$$

TUNING

X=(a, 1)

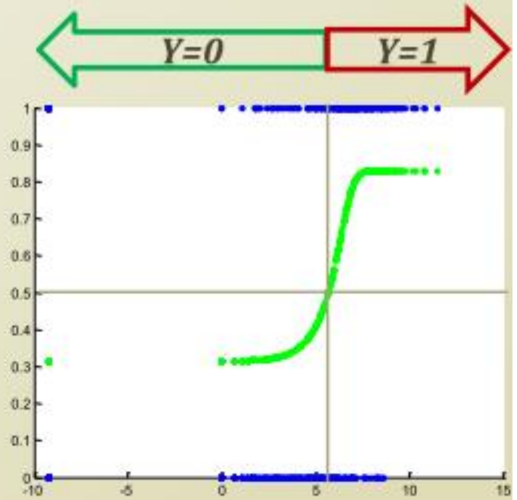
Logit→Logistic
Inverse of X and Y
X in Logit is the probability

Linear shift for a better function fitting

- When we are fitting the linear regression to approximate P(Y|X)
  - $X\theta = P(Y|X)$
  - Though, this is not going to keep the probability axiom
- Now we are fitting to the logistic function to approximate P(Y|X)
  - $X\theta = \log\left(\frac{P(Y|X)}{1-P(Y|X)}\right)$
  - From linear to logistic

The Logistic regression classifier

Logistic Regression



- Logistic regression is a probabilistic classifier to predict the binomial or the multinomial outcome
  - by fitting the conditional probability to the logistic function.
- You can see the problem from the different view.
  - This way is actually closer to the formal definition.
- Given the Bernoulli experiment
  - $P(y|x) = \mu(x)^y(1-\mu(x))^{1-y}$

    accident
  - $\mu(x) = \frac{1}{1+e^{-\theta^T x}} = P(y=1|x)$
  - Here, $\mu(x)$ is the logistic function
- From the previous slide,
  - $X\theta = \log\left(\frac{P(Y|X)}{1-P(Y|X)}\right) \rightarrow P(Y|X) = \boxed{\frac{e^{X\theta}}{1+e^{X\theta}}}$

Logistic Function

$$f(x) = \frac{1}{1+e^{-x}}$$

The goal, finally, becomes finding out $\boldsymbol{\theta}$, again

Related Patent(KR; 독립항에 로지스틱 회귀 포함)

No	상태	문헌번호	문헌일	출원일	만료(예상)일	발명의 명칭	출원인
1	등록	KR 10-1573368 B1	2015.11.25	2014.08.08	2034.08.08	로지스틱 대장균 반응수와 퍼지 수질 소속함수를 합성하여 하천 레크리에이션 지수를 산정하는 방법 (Method for estimating River Recreational Index integrating logistic fecal coliform function and fuzzy water quality membership function)	서울대학교산학협력단
2	등록	KR 10-1535189 B1	2015.07.02	2015.05.14	2035.05.14	로지스틱 회귀분석법을 이용한 스폿용접 품질평가 시스템 및 그의 방법 (System of spot welding quality evaluation using logistic regression and Method the same)	한양대학교 산학협력단
3	등록	KR 10-1493552 B1	2015.02.09	2008.05.14	2028.05.14	신호판정방법, 신호판정장치, 프로그램, 신호판정시스템 (Signal judgment method, signal judgment apparatus, program, and signal judgment system)	넛토보 온쿄 엔지니어링 가부시키가이샤
4	등록	KR 10-1378238 B1	2014.03.19	2010.09.30	2030.09.30	전통 중의학 (TCM) 원리에 기초한 피부 조성을 결정하기 위한 컴퓨터 이용 진단 시스템 및 방법 (COMPUTER-AIDED DIAGNOSTIC SYSTEMS AND METHODS FOR DETERMINING SKIN COMPOSITIONS BASED ON TRADITIONAL CHINESE MEDICINAL(TCM) PRINCIPLES)	이엘씨 매니지먼트 엘엘씨
5	등록	KR 10-1278693 B1	2013.06.17	2011.10.14	2031.10.14	합격률 예측 방법 및 그 장치 (METHOD AND APPARATUS FOR PREDICTING PASS PROBABILITY OF EXAMINATION)	주식회사 에듀스 | 박수한
6	등록	KR 10-1268766 B1	2013.05.22	2011.01.20	2031.01.20	중증 천식의 악화 진단용 기상 및 대기 오염 인자의 위험도 예측방법 (METHOD FOR PREDICTING RISK OF METEOROLOGICAL FACTORS AND AIR POLLUTION FACTORS FOR DIAGNOSING EXACERBATION OF REFRACTORY ASTHMA)	순천향대학교 산학협력단
7	등록	KR 10-1102496 B1	2011.12.28	2010.10.18	2030.10.18	전력 기자재 상태판정식 도출 방법, 이를 이용한 활선상태의 전력 기자재 진단 장치 및 방법 (Method for deriving state determination equation, Apparatus and Method for diagnosis of power equipment using the same)	(주)이아이에스글로벌 | 한국전력공사
8	등록	KR 10-0952681 B1	2010.04.06	2009.06.22	2029.06.22	유전자 분석을 이용한 천식 또는 천식 발병 가능성 판별 방법 및 장치 (Methods and devices for detection of asthma or risk thereof using genetic analysis)	대전대학교 산학협력단
9	등록	KR 10-0590547 B1	2006.06.09	2004.02.28	2024.02.28	복합 질환과 연관된 다중 SNP 마커들로부터 최적 마커세트를 선택하는 방법 (A method for selecting optimized SNP marker sets associated with a complex disease from multiple SNP markers)	삼성전자주식회사

적용 분야가 통계 처리 관련된 것으로 제한적

Related Patent(KR; 독립항에 로지스틱 회귀 포함)

청구항 1항 (대표청구항)

컴퓨터에 의해 소음분석 대상을 판정하는 신호판정방법으로서,
소음 또는 진동에 관계되는 측정분석 대상의 모델에, 측정분석 대상인지가 미지인 미지 데이터를 입력하고,
상기 측정분석 대상의 모델의 출력값을, 상기 미지 데이터가 측정분석 대상인지의 확률값으로서 얻는
것을 특징으로 하는 신호판정방법으로서,
상기 측정분석 대상의 모델은, 실측된 기지 데이터를 사용하여 제작하고,
상기 실측된 기지 데이터를 사용하여 상기 측정분석 대상의 모델을 예측 모델식 산출수단에 의해 제작하고,
상기 측정분석 대상의 모델을 사용한 출력값 산출수단에 의한 계산의 출력값을, 상기 미지 데이터가 측정분석 대상인지의 확률값으로서 얻고,
상기 측정분석 대상의 모델에 관계되는 예측 인자를 설명 변수로서 사용하는
것을 특징으로 하는 신호판정방법.

청구항 4항

제1항에 있어서,
상기 측정분석 대상의 모델에 관계되는 예측 모델식에, 로지스틱 회귀식을 사용하고,
상기 로지스틱 회귀식은,
상기 측정분석 대상인지 여부를 목적 변수로서 사용하고,
상기 로지스틱 회귀식에, 상기 미지 데이터를 적용해서 입력하여, 확률값을 산출하는,
것을 특징으로 하는 신호판정방법.

컴퓨터 상에서 일별 기상 및 대기 오염 인자 중 천식 악화에 영향을 미치는 인자를 선정하는 단계;
조건부 로지스틱 회귀분석(Conditional logistic regression) 방법을 사용하여 컴퓨터 상에서 상기 선정한 인자들을 모델링하는 단계; 및
케이스-크로스오버(case-crossover)방법을 사용하여 컴퓨터 상에서 상기 선정한 인자들을 정량 분석하는 단계를 포함하는 중증 천식(Refractory asthma)의 악화 진단용 기상 및 대기 오염 인자의 위험도 예측방법으로,
상기 케이스-크로스오버 방법은,
중증 천식의 악화 당일(T=0) 내지 3일 전(T=3)의 위험도와 대조군의 위험도를 비교하여 위험도 계산식 및 위험도 예측 인자를 선정하는 것을 특징으로 하는 중증 천식의 악화 진단용 기상 및 대기 오염 인자의 위험도 예측방법.

Related patents

(12) **United States Patent**
Corrado et al.

(54) **USING EMBEDDING FUNCTIONS WITH A DEEP NETWORK**

(71) Applicant: **Google Inc.**, Mountain View, CA (US)

(72) Inventors: **Gregory S. Corrado**, San Francisco, CA (US); **Kai Chen**, Brisbane, CA (US); **Jeffrey A. Dean**, Palo Alto, CA (US); **Gary R. Holt**, Murrysville, PA (US); **Julian P. Grady**, Pittsburgh, PA (US); **Sharat Chikkerur**, Pittsburgh, PA (US); **David W. Sculley**, Pittsburgh, PA (US)

(73) Assignee: **Google Inc.**, Mountain View, CA (US)

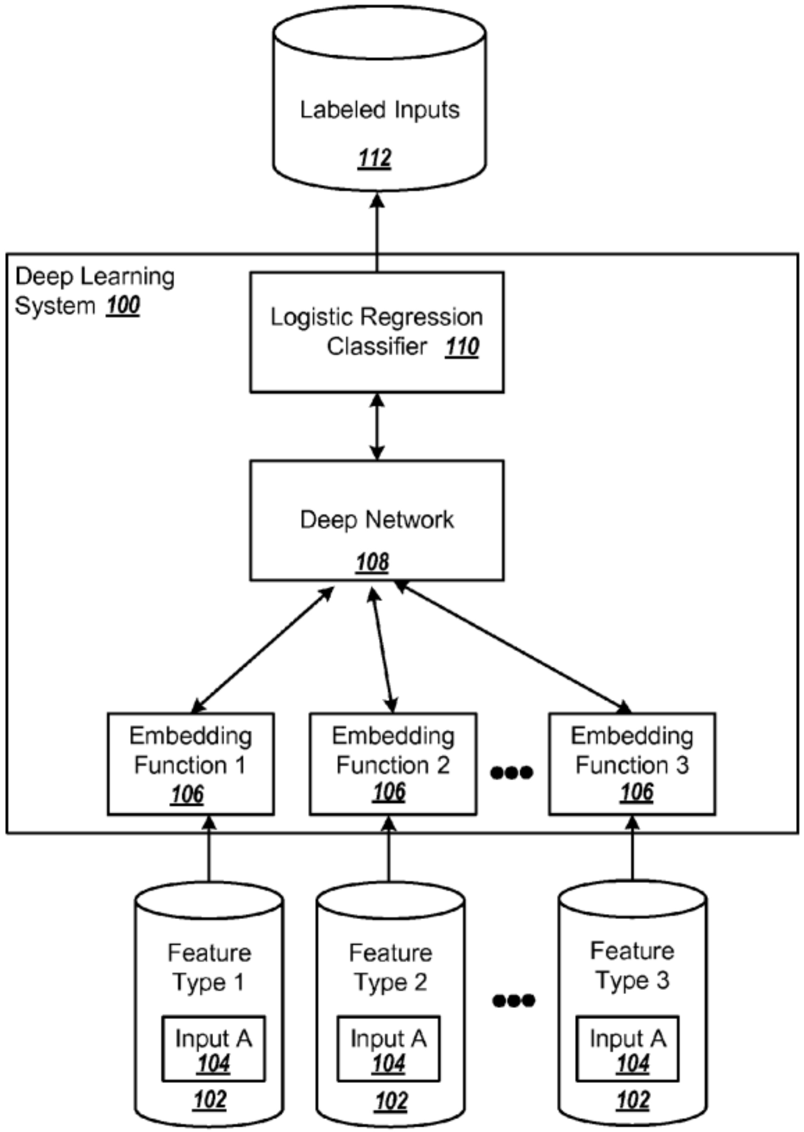
(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 272 days.

(21) Appl. No.: **13/803,779**

(22) Filed: **Mar. 14, 2013**

Related U.S. Application Data

(60) Provisional application No. 61/666,684, filed on Jun. 29, 2012.



Related Patents

1. A method performed by one or more computers, the method comprising:

receiving an input comprising a plurality of features, wherein each of the features is of a different feature type;

processing each of the features using a respective embedding function to generate one or more respective numeric values, wherein each of the embedding functions operates independently of each of the other embedding functions, and wherein each of the embedding functions is specific to features of a respective feature type;

processing the numeric values using a deep network to generate a first alternative representation of the input, wherein the deep network is a machine learning model composed of a plurality of levels of non-linear operations; and

processing the first alternative representation of the input using a logistic regression classifier to predict a label for the input.

Related Patents(litigation related)

미국 등록 특허 중, 소송과 관련되어 있고, 독립항에 'regression'이 기재된 특허 list

☐ 1	US8332822B2 2012-12-11 Technologies for code failure proneness estimation
☐ 2	US7110585B2 2006-09-19 Nanoparticle imaging system and method
☐ 3	US6892113B1 2005-05-10 Irrigation controller using regression model
☐ 4	US6804625B1 2004-10-12 Subsurface modeling method
☐ 5	US6681787B2 2004-01-27 System and method of operation of a digital mass flow controller
☐ 6	US6470279B1 2002-10-22 Method for calibrating spectrophotometric apparatus with synthetic fluids to measure plasma and serum analytes
☐ 7	US6427141B1 2002-07-30 Enhancing knowledge discovery using multiple support vector machines
☐ 8	US6192998B1 2001-02-27 Method of and system for optimizing rate of penetration in drilling operations
☐ 9	US6155357A 2000-12-05 Method of and system for optimizing rate of penetration in drilling operations
☐ 10	US5987399A 1999-11-16 Ultrasensitive surveillance of sensors and processes
☐ 11	US5848396A 1998-12-08 Method and apparatus for determining behavioral profile of a computer user
☐ 12	US5670537A 1997-09-23 Method for effecting tumor regression with a low dose, short infusion taxol regimen

Related Patents(litigation related)

Data Engine Technologies LLC v. Microsoft Corporation, Filed October 1, 2013, **D.C. E.D. Texas, Doc. No.** 6:13cv735

(12) **United States Patent**
Nagappan et al.

(10) **Patent No.: US 8,332,822 B2**
(45) **Date of Patent: *Dec. 11, 2012**

(54) **TECHNOLOGIES FOR CODE FAILURE PRONENESS ESTIMATION**

(75) Inventors: **Nachiappan Nagappan**, Redmond, WA (US); **Thirumalesh Bhat**, Sammamish, WA (US)

(73) Assignee: **Microsoft Corporation**, Redmond, WA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

This patent is subject to a terminal disclaimer.

(21) Appl. No.: **13/042,404**

(22) Filed: **Mar. 7, 2011**

(65) **Prior Publication Data**

US 2011/0161932 A1 Jun. 30, 2011

Related U.S. Application Data

(63) Continuation of application No. 11/740,722, filed on Apr. 26, 2007, now Pat. No. 7,926,036.

1. A method comprising:
calculating, by a computing device, churn metrics that indicate a degree of change between a first version of code and a second version of the code;
calculating, by the computing device for the second version and based on input code metrics that include the churn metrics, second metrics determined from the second version, and first metrics that correspond to the churn metrics and the second metrics that are determined from the first version, a historical variant metric feedback factor that is a sum divided by a total number of the input code metrics, where the sum is a count of each of the input code metrics that exceed a statistical upper bound that is based on the each of the input code metrics' corresponding first metric; and
performing, by the computing device and based on the second metrics or on the churn metrics, a logistical regression resulting in a code failure proneness probability for the second version.

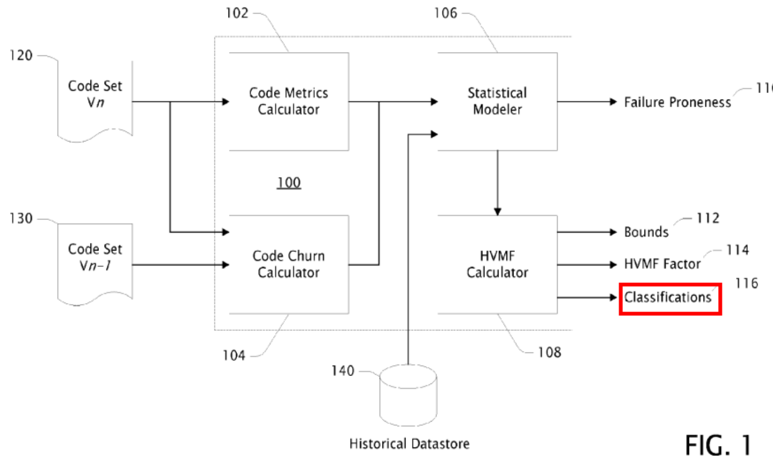


FIG. 1

Related Patents(litigation related)

Docket Number	Docket Description	Court	Date Filed
2:15cv371	Smart Irrigation Solutions V. The Toro Company	UNITED STATES DISTRICT COURT for the CENTRAL DISTRICT OF CALIFORNIA	1/16/2015
8:15cv81	Smart Irrigation Solutions Inc. V. Hunter Industries Incorporated	UNITED STATES DISTRICT COURT for the CENTRAL DISTRICT OF CALIFORNIA	1/16/2015
8:15cv84	Smart Irrigation Solutions V. The Toro Company	UNITED STATES DISTRICT COURT for the CENTRAL DISTRICT OF CALIFORNIA	1/16/2015

ATTACK!

(12) **United States Patent**
Addink et al.

(10) **Patent No.: US 6,892,113 B1**
(45) **Date of Patent: May 10, 2005**

(54) **IRRIGATION CONTROLLER USING REGRESSION MODEL**

(75) Inventors: **John Addink**, Riverside, CA (US); **Sylvan Addink**, Iowa City, CA (US)

(73) Assignee: **Aqua Conserve, Inc.**, Riverside, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 0 days.

(21) Appl. No.: **10/009,867**

(22) PCT Filed: **Jul. 7, 2000**

(86) PCT No.: **PCT/US00/18705**

§ 371 (c)(1), (2), (4) Date: **Dec. 11, 2001**

(87) PCT Pub. No.: **WO02/05045**

PCT Pub. Date: **Jan. 17, 2002**

(57) **ABSTRACT**

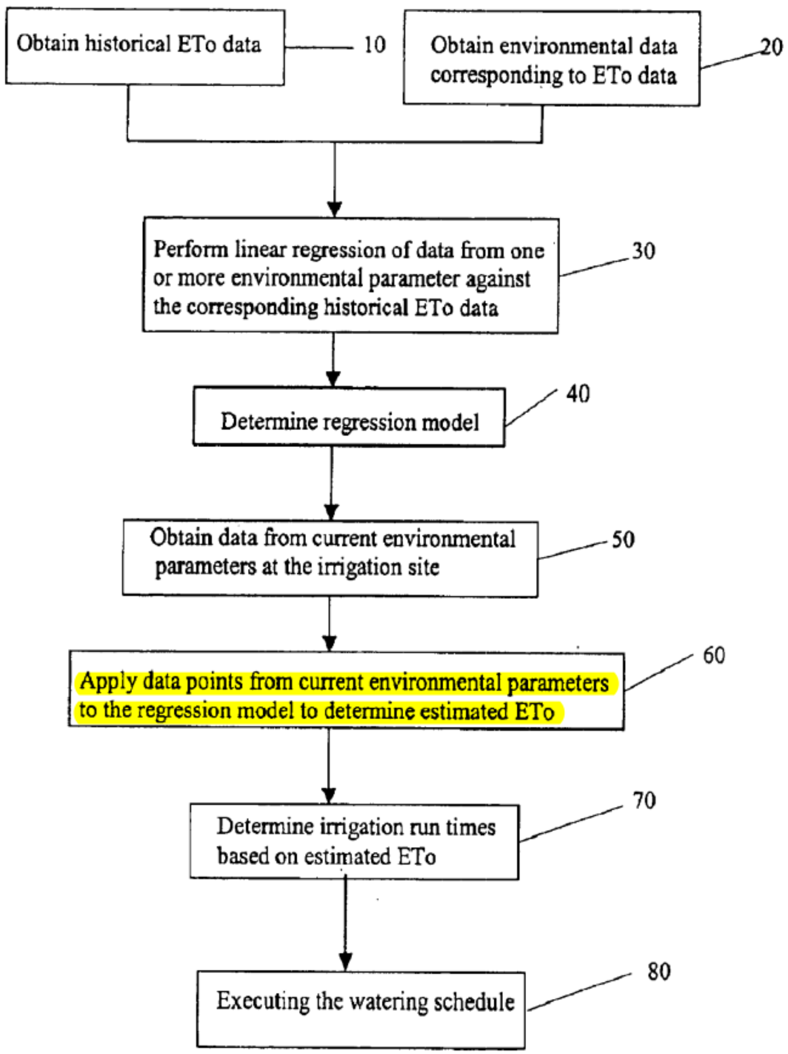
The present invention provides systems and methods in which an irrigation controller uses a regression model to estimate an evapotranspiration rate (estimated ETo), and uses the estimated ETo to affect an irrigation schedule executed by the controller. The regression model is preferably based upon a comparison of historical ETo values against corresponding historical environmental values, with the data advantageously spanning a time period of at least one month, and more preferably at least two months. Data for multiple environmental factors may also be used. The environmental factor(s) utilized may advantageously comprise one or more of temperature, solar radiation, wind speed, humidity, barometric pressure, and soil moisture. Values relating the environmental factor(s) may enter the controller from a local sensor, a distal signal source, or both.

claim

1. An irrigation controller comprising:

a memory that stores a regression model, wherein the regression model is based upon a set of historical ETo values and a set of corresponding historical values for an environmental factor, the regression model running with optional input from a local sensor;

a microprocessor that applies a current value for the environmental factor to the regression model to calculate a current evapotranspiration rate (estimated ETo), wherein the irrigation controller uses the estimated ETo to determine an irrigation schedule executed by the irrigation controller.



감사합니다.

참고 문헌

1. KAIST SESLAB 강의 자료 - http://seslab.kaist.ac.kr/xe2/page_GBex27

2. https://azure.microsoft.com/en-us/documentation/articles/machine-learning-algorithm-choice/