Title: Generative model

URL: https://en.wikipedia.org/wiki/Generative_model

PageID: 1222578

Categories: Category:Machine learning, Category:Probabilistic models, Category:Statistical models

-----

In statistical classification , two main approaches are called the generative approach and the discriminative approach. These compute classifiers by different approaches, differing in the degree of statistical modelling . Terminology is inconsistent, but three major types can be distinguished:

A generative model is a statistical model of the joint probability distribution P (X , Y) {\displaystyle P(X,Y)} on a given observable variable X and target variable Y ; A generative model can be used to "generate" random instances (outcomes) of an observation x .

A discriminative model is a model of the conditional probability P (Y ■ X = x) {\displaystyle P(Y\mid X=x)} of the target Y , given an observation x . It can be used to "discriminate" the value of the target variable Y , given an observation x .

Classifiers computed without using a probability model are also referred to loosely as "discriminative".

The distinction between these last two classes is not consistently made; Jebara (2004) refers to these three classes as generative learning , conditional learning , and discriminative learning , but Ng & Jordan (2002) only distinguish two classes, calling them generative classifiers (joint distribution) and discriminative classifiers (conditional distribution or no distribution), not distinguishing between the latter two classes. Analogously, a classifier based on a generative model is a generative classifier, while a classifier based on a discriminative model is a discriminative classifier, though this term also refers to classifiers that are not based on a model.

Standard examples of each, all of which are linear classifiers , are:

generative classifiers: naive Bayes classifier and linear discriminant analysis

naive Bayes classifier and

linear discriminant analysis

discriminative model: logistic regression

logistic regression

In application to classification, one wishes to go from an observation x to a label y (or probability distribution on labels). One can compute this directly, without using a probability distribution (distribution-free classifier); one can estimate the probability of a label given an observation, P (Y | X = x) {\displaystyle P(Y|X=x)} (discriminative model), and base classification on that; or one can estimate the joint distribution P (X , Y) {\displaystyle P(X,Y)} (generative model), from that compute the conditional probability P (Y | X = x) {\displaystyle P(Y|X=x)} , and then base classification on that. These are increasingly indirect, but increasingly probabilistic, allowing more domain knowledge and probability theory to be applied. In practice different approaches are used, depending on the particular problem, and hybrids can combine strengths of multiple approaches.

Definition

An alternative division defines these symmetrically as:

a generative model is a model of the conditional probability of the observable X , given a target y , symbolically, P (X ■ Y = y) {\displaystyle P(X\mid Y=y)}

a discriminative model is a model of the conditional probability of the target Y , given an observation x , symbolically, P (Y ■ X = x) {\displaystyle P(Y\mid X=x)}

Regardless of precise definition, the terminology is constitutional because a generative model can be used to "generate" random instances (outcomes), either of an observation and target (x , y) {\displaystyle (x,y)} , or of an observation x given a target value y , while a discriminative model or discriminative classifier (without a model) can be used to "discriminate" the value of the target variable Y , given an observation x . The difference between " discriminate " (distinguish) and " classify " is subtle, and these are not consistently distinguished. (The term "discriminative classifier" becomes a pleonasm when "discrimination" is equivalent to "classification".)

The term "generative model" is also used to describe models that generate instances of output variables in a way that has no clear relationship to probability distributions over potential samples of input variables. Generative adversarial networks are examples of this class of generative models, and are judged primarily by the similarity of particular outputs to potential inputs. Such models are not classifiers.

Relationships between models

In application to classification, the observable X is frequently a continuous variable , the target Y is generally a discrete variable consisting of a finite set of labels, and the conditional probability P (Y ■ X) {\displaystyle P(Y\mid X)} can also be interpreted as a (non-deterministic) target function f : X → Y {\displaystyle f\colon X\to Y} , considering X as inputs and Y as outputs.

Given a finite set of labels, the two definitions of "generative model" are closely related. A model of the conditional distribution P (X ■ Y = y) {\displaystyle P(X\mid Y=y)} is a model of the distribution of each label, and a model of the joint distribution is equivalent to a model of the distribution of label values P (Y) {\displaystyle P(Y)} , together with the distribution of observations given a label, P (X ■ Y) {\displaystyle P(X\mid Y)} ; symbolically, P (X , Y) = P (X ■ Y) P (Y) . {\displaystyle P(X,Y)=P(X\mid Y)P(Y).} Thus, while a model of the joint probability distribution is more informative than a model of the distribution of label (but without their relative frequencies), it is a relatively small step, hence these are not always distinguished.

Given a model of the joint distribution, P (X , Y) {\displaystyle P(X,Y)} , the distribution of the individual variables can be computed as the marginal distributions P (X) = ∑ y P (X , Y = y) {\displaystyle P(X)=\sum _{y}P(X,Y=y)} and P (Y) = ∫ x P (Y , X = x) {\displaystyle P(Y)=\int _{x}P(Y,X=x)} (considering X as continuous, hence integrating over it, and Y as discrete, hence summing over it), and either conditional distribution can be computed from the definition of conditional probability : P (X ■ Y) = P (X , Y) / P (Y) {\displaystyle P(X\mid Y)=P(X,Y)/P(Y)} and P (Y ■ X) = P (X , Y) / P (X) {\displaystyle P(Y\mid X)=P(X,Y)/P(X)} .

Given a model of one conditional probability, and estimated probability distributions for the variables X and Y , denoted P (X) {\displaystyle P(X)} and P (Y) {\displaystyle P(Y)} , one can estimate the opposite conditional probability using Bayes' rule :

For example, given a generative model for P (X ■ Y) {\displaystyle P(X\mid Y)} , one can estimate:

and given a discriminative model for P (Y ■ X) {\displaystyle P(Y\mid X)} , one can estimate:

Note that Bayes' rule (computing one conditional probability in terms of the other) and the definition of conditional probability (computing conditional probability in terms of the joint distribution) are frequently conflated as well.

Contrast with discriminative classifiers

A generative algorithm models how the data was generated in order to categorize a signal. It asks the question: based on my generation assumptions, which category is most likely to generate this signal? A discriminative algorithm does not care about how the data was generated, it simply categorizes a given signal. So, discriminative algorithms try to learn p (y | x) {\displaystyle p(y|x)} directly from the data and then try to classify data. On the other hand, generative algorithms try to learn p (x , y) {\displaystyle p(x,y)} which can be transformed into p (y | x) {\displaystyle p(y|x)} later to classify the data. One of the advantages of generative algorithms is that you can use p (x , y) {\displaystyle p(x,y)} to generate new data similar to existing data. On the other hand, it has been proved that some discriminative algorithms give better performance than some generative algorithms in classification tasks.

Despite the fact that discriminative models do not need to model the distribution of the observed variables, they cannot generally express complex relationships between the observed and target variables. But in general, they don't necessarily perform better than generative models at classification and regression tasks. The two classes are seen as complementary or as different views of the same procedure.

Deep generative models

With the rise of deep learning , a new family of methods, called deep generative models (DGMs), is formed through the combination of generative models and deep neural networks. An increase in the scale of the neural networks is typically accompanied by an increase in the scale of the training data, both of which are required for good performance.

Popular DGMs include variational autoencoders (VAEs), generative adversarial networks (GANs), and auto-regressive models. Recently, there has been a trend to build very large deep generative models. For example, GPT-3 , and its precursor GPT-2 , are auto-regressive neural language models that contain billions of parameters, BigGAN and VQ-VAE which are used for image generation that can have hundreds of millions of parameters, and Jukebox is a very large generative model for musical audio that contains billions of parameters.

Types

Generative models

Types of generative models are:

Gaussian mixture model (and other types of mixture model)

Hidden Markov model

Probabilistic context-free grammar

Bayesian network (e.g. Naive bayes , Autoregressive model)

Averaged one-dependence estimators

Latent Dirichlet allocation

Boltzmann machine (e.g. Restricted Boltzmann machine , Deep belief network)

Variational autoencoder

Generative adversarial network

Flow-based generative model

Energy based model

Diffusion model

If the observed data are truly sampled from the generative model, then fitting the parameters of the generative model to maximize the data likelihood is a common method. However, since most statistical models are only approximations to the true distribution, if the model's application is to infer about a subset of variables conditional on known values of others, then it can be argued that the approximation makes more assumptions than are necessary to solve the problem at hand. In such cases, it can be more accurate to model the conditional density functions directly using a discriminative model (see below), although application-specific details will ultimately dictate which approach is most suitable in any particular case.

Discriminative models

k-nearest neighbors algorithm

Logistic regression

Support Vector Machines

Decision Tree Learning

Random Forest

Maximum-entropy Markov models

Conditional random fields

Examples

Simple example

Suppose the input data is x ∈ { 1 , 2 } {\displaystyle x\in \{1,2\}} , the set of labels for x {\displaystyle x} is y ∈ { 0 , 1 } {\displaystyle y\in \{0,1\}} , and there are the following 4 data points: (x , y) = { (1 , 0) , (1 , 1) , (2 , 0) , (2 , 1) } {\displaystyle (x,y)=\{(1,0),(1,1),(2,0),(2,1)\}}

For the above data, estimating the joint probability distribution p (x , y) {\displaystyle p(x,y)} from the empirical measure will be the following:

while p (y | x) {\displaystyle p(y|x)} will be following:

Text generation

Shannon (1948) gives an example in which a table of frequencies of English word pairs is used to generate a sentence beginning with "representing and speedily is an good"; which is not proper English but which will increasingly approximate it as the table is moved from word pairs to word triplets etc.

See also

Mathematics portal

Discriminative model

Graphical model

Notes

References

External links

Shannon, C. E. (1948). "A Mathematical Theory of Communication" (PDF) . Bell System Technical Journal . 27 (July, October): 379– 423, 623– 656. doi : 10.1002/j.1538-7305.1948.tb01338.x . hdl : 10338.dmlcz/101429 . Archived from the original (PDF) on 2016-06-06 . Retrieved 2016-01-09 .

Mitchell, Tom M. (2015). "3. Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression" (PDF) . Machine Learning .

Ng, Andrew Y. ; Jordan, Michael I. (2002). "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes" (PDF) . Advances in Neural Information Processing Systems .

Jebara, Tony (2002). Discriminative, generative, and imitative learning (PhD). Massachusetts Institute of Technology . hdl : 1721.1/8323 . , (mirror , mirror), published as book (above)

v

t

e

Outline

Index

Mean Arithmetic Arithmetic-Geometric Contraharmonic Cubic Generalized/power Geometric Harmonic Heronian Heinz Lehmer

Arithmetic

Arithmetic-Geometric

Contraharmonic

Cubic

Generalized/power

Geometric

Harmonic

Heronian

Heinz

Lehmer

Median

Mode

Average absolute deviation

Coefficient of variation

Interquartile range

Percentile

Range

Standard deviation

Variance

Central limit theorem

Moments Kurtosis L-moments Skewness

Kurtosis

L-moments

Skewness

Index of dispersion

Contingency table

Frequency distribution

Grouped data

Partial correlation

Pearson product-moment correlation

Rank correlation Kendall's τ Spearman's ρ

Kendall's τ

Spearman's ρ

Scatter plot

Bar chart

Biplot

Box plot

Control chart

Correlogram

Fan chart

Forest plot

Histogram

Pie chart

Q–Q plot

Radar chart

Run chart

Scatter plot

Stem-and-leaf display

Violin plot

Effect size

Missing data

Optimal design

Population

Replication

Sample size determination

Statistic

Statistical power

Sampling Cluster Stratified

Cluster

Stratified

Opinion poll

Questionnaire

Standard error

Blocking

Factorial experiment

Interaction

Random assignment

Randomized controlled trial

Randomized experiment

Scientific control

Adaptive clinical trial

Stochastic approximation

Up-and-down designs

Cohort study

Cross-sectional study

Natural experiment

Quasi-experiment

Population

Statistic

Probability distribution

Sampling distribution Order statistic

Order statistic

Empirical distribution Density estimation

Density estimation

Statistical model Model specification L p space

Model specification

L p space

Parameter location scale shape

location

scale

shape

Parametric family Likelihood (monotone) Location–scale family Exponential family

Likelihood (monotone)

Location–scale family

Exponential family

Completeness

Sufficiency

Statistical functional Bootstrap U V

Bootstrap

U

V

Optimal decision loss function

loss function

Efficiency

Statistical distance divergence

divergence

Asymptotics

Robustness

Estimating equations Maximum likelihood Method of moments M-estimator Minimum distance

Maximum likelihood

Method of moments

M-estimator

Minimum distance

Unbiased estimators Mean-unbiased minimum-variance Rao–Blackwellization Lehmann–Scheffé theorem Median unbiased

Mean-unbiased minimum-variance Rao–Blackwellization Lehmann–Scheffé theorem

Rao–Blackwellization

Lehmann–Scheffé theorem

Median unbiased

Plug-in

Confidence interval

Pivot

Likelihood interval

Prediction interval

Tolerance interval

Resampling Bootstrap Jackknife

Bootstrap

Jackknife

1- & 2-tails

Power Uniformly most powerful test

Uniformly most powerful test

Permutation test Randomization test

Randomization test

Multiple comparisons

Likelihood-ratio

Score/Lagrange multiplier

Wald

Z -test (normal)

Student's t -test

F -test

Chi-squared

G -test

Kolmogorov–Smirnov

Anderson–Darling

Lilliefors

Jarque–Bera

Normality (Shapiro–Wilk)

Likelihood-ratio test

Model selection Cross validation AIC BIC

Cross validation

AIC

BIC

Sign Sample median

Sample median

Signed rank (Wilcoxon) Hodges–Lehmann estimator

Hodges–Lehmann estimator

Rank sum (Mann–Whitney)

Nonparametric anova 1-way (Kruskal–Wallis) 2-way (Friedman) Ordered alternative (Jonckheere–Terpstra)

1-way (Kruskal–Wallis)

2-way (Friedman)

Ordered alternative (Jonckheere–Terpstra)

Van der Waerden test

Bayesian probability prior posterior

prior

posterior

Credible interval

Bayes factor

Bayesian estimator Maximum posterior estimator

Maximum posterior estimator

Correlation

Regression analysis

Pearson product-moment

Partial correlation

Confounding variable

Coefficient of determination

Errors and residuals

Regression validation

Mixed effects models

Simultaneous equations models

Multivariate adaptive regression splines (MARS)

Simple linear regression

Ordinary least squares

General linear model

Bayesian regression

Nonlinear regression

Nonparametric

Semiparametric

Isotonic

Robust

Homoscedasticity and Heteroscedasticity

Exponential families

Logistic (Bernoulli) / Binomial / Poisson regressions

Analysis of variance (ANOVA, anova)

Analysis of covariance

Multivariate ANOVA

Degrees of freedom

Cohen's kappa

Contingency table

Graphical model

Log-linear model

McNemar's test

Cochran–Mantel–Haenszel statistics

Regression

Manova

Principal components

Canonical correlation

Discriminant analysis

Cluster analysis

Classification

Structural equation model Factor analysis

Factor analysis

Multivariate distributions Elliptical distributions Normal

Elliptical distributions Normal

Normal

Decomposition

Trend

Stationarity

Seasonal adjustment

Exponential smoothing

Cointegration

Structural break

Granger causality

Dickey–Fuller

Johansen

Q-statistic (Ljung–Box)

Durbin–Watson

Breusch–Godfrey

Autocorrelation (ACF) partial (PACF)

partial (PACF)

Cross-correlation (XCF)

ARMA model

ARIMA model (Box–Jenkins)

Autoregressive conditional heteroskedasticity (ARCH)

Vector autoregression (VAR) (Autoregressive model (AR))

Spectral density estimation

Fourier analysis

Least-squares spectral analysis

Wavelet

Whittle likelihood

Kaplan–Meier estimator (product limit)

Proportional hazards models

Accelerated failure time (AFT) model

First hitting time

Nelson–Aalen estimator

Log-rank test

Bioinformatics

Clinical trials / studies

Epidemiology

Medical statistics

Chemometrics

Methods engineering

Probabilistic design

Process / quality control

Reliability

System identification

Actuarial science

Census

Crime statistics

Demography

Econometrics

Jurimetrics

National accounts

Official statistics

Population statistics

Psychometrics

Cartography

Environmental statistics

Geographic information system

Geostatistics

Kriging

Category

Mathematics portal

Commons

WikiProject