WIKIPEDIA

Generative model

In statistical classification, including machine learning, 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,[a] but three major types can be distinguished, following Jebara (2004):

- Given an observable variable X and a target variable Y, a **generative model** is a statistical model of the joint probability distribution on $X \times Y$, $P(X,Y)$;[1]
- A **discriminative model** is a model of the conditional probability of the target Y, given an observation x, symbolically, $P(Y|X=x)$; and
- 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;[2] 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.[3] 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
- discriminative model:
  - logistic regression
  - non-model classifier:
    - perceptron and
    - support vector machine.

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, $P(Y|X=x)$ (*discriminative model*), and base classification on that; or one can estimate the joint distribution $P(X,Y)$ (*generative model*), from that compute the conditional probability $P(Y|X=x)$, 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.

Contents

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, $P(X|Y=y)$[4]
- a **discriminative model** is a model of the conditional probability of the target *Y*, given an observation *x*, symbolically, $P(Y|X=x)$[5]

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), or of an observation x given a target value y,[4] 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.[5] 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 $P(Y|X)$ can also be interpreted as a (non-deterministic) target function $f\colon X \to Y$, 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 $P(X|Y=y)$ 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 $P(Y)$, together with the distribution of observations given a label, $P(X|Y)$; symbolically, $P(X,Y)=P(X|Y)P(Y)$. 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, $P(X,Y)$, the distribution of the individual variables can be computed as the marginal distributions $P(X)=\sum_y P(X,Y=y)$ and $P(Y)=\int_x P(Y,X=x)$ (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: $P(X|Y)=P(X,Y)/P(Y)$ and $P(Y|X)=P(X,Y)/P(X)$.

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

$$P(X|Y)P(Y)=P(Y|X)P(X).$$

For example, given a generative model for $P(X|Y)$, one can estimate:

$$P(Y|X)=P(X|Y)P(Y)/P(X),$$

and given a discriminative model for $P(Y|X)$, one can estimate:

$$P(X|Y)=P(Y|X)P(X)/P(Y).$$

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 $p(y|x)$ directly from the data and then try to classify data. On the other hand, generative algorithms try to learn $p(x,y)$ which can be transformed into $p(y|x)$ later to classify the data. One of the advantages of generative algorithms is that you can use $p(x,y)$ to generate new data similar to existing data. On the other hand, discriminative algorithms generally give better performance in classification tasks.[6]

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. 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.[7]

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

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
- Maximum-entropy Markov models
- Conditional random fields
- Neural networks

Examples

Simple example

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

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

	$y=0$	$y=1$
$x=1$	$1/2$	0
$x=2$	$1/4$	$1/4$

while $p(y|x)$ will be following:

	$y=0$	$y=1$
$x=1$	1	0
$x=2$	$1/2$	$1/2$

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

- Discriminative model
- Graphical model

Notes

a. Three leading sources, Ng & Jordan 2002, Jebara 2004, and Mitchell 2015, give different divisions and definitions.

References

1. Ng & Jordan (2002): "Generative classifiers learn a model of the joint probability, $p(x,y)$, of the inputs *x* and the label *y*, and make their predictions by using Bayes rules to calculate $p(y|x)$, and then picking the most likely label *y*.
2. Jebara 2004, 2.4 Discriminative Learning: "This distinction between conditional learning and discriminative learning is not currently a well established convention in the field."
3. Ng & Jordan 2002: "Discriminative classifiers model the posterior $p(y|x)$ directly, or learn a direct map from inputs *x* to the class labels."
4. Mitchell 2015: "We can use Bayes rule as the basis for designing learning algorithms (function approximators), as follows: Given that we wish to learn some target function $f\colon X\to Y$, or equivalently, $P(Y|X)$, we use the training data to learn estimates of $P(X|Y)$ and $P(Y)$. New *X* examples can then be classified using these estimated probability distributions, plus Bayes rule. This type of classifier is called a *generative* classifier, because we can view the distribution $P(X|Y)$ as describing how to generate random instances *X* conditioned on the target attribute *Y*.
5. Mitchell 2015: "Logistic Regression is a function approximation algorithm that uses training data to directly estimate $P(Y|X)$, in contrast to Naive Bayes. In this sense, Logistic Regression is often referred to as a *discriminative* classifier because we can view

the distribution $P(Y|X)$ as directly discriminating the value of the target value Y for any given instance X

6. Ng & Jordan 2002
7. Bishop, C. M.; Lasserre, J. (24 September 2007), "Generative or Discriminative? getting the best of both worlds", in Bernardo, J. M. (ed.), *Bayesian statistics 8: proceedings of the eighth Valencia International Meeting, June 2-6, 2006* (https://books.google.com/books?id=Vh7vAAAAMAAJ&pg=PA3), Oxford University Press, pp. 3–23, ISBN 978-0-19-921465-5

External links

- Shannon, C. E. (1948). "A Mathematical Theory of Communication" (https://www.tnt.uni-hannover.de/edu/vorlesungen/InfoTheor/download/shannon1948.pdf) (PDF). *Bell System Technical Journal*. **27** (July, October): 379–423, 623–656. doi:10.1002/j.1538-7305.1948.tb01338.x (https://doi.org/10.1002%2Fj.1538-7305.1948.tb01338.x). hdl:10338.dmlcz/101429 (https://hdl.handle.net/10338.dmlcz%2F101429).
- Mitchell, Tom M. (2015). "3. Generative and Discriminative Classifiers: Naive Bayes and Logistic Regression" (https://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf) (PDF). *Machine Learning*.
- Ng, Andrew Y.; Jordan, Michael I. (2002). "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes" (http://robotics.stanford.edu/~ang/papers/nips01-discriminativegenerative.pdf) (PDF). *Advances in Neural Information Processing Systems*.
- Jebara, Tony (2004). *Machine Learning: Discriminative and Generative* (https://www.springer.com/us/book/9781402076473). The Springer International Series in Engineering and Computer Science. Kluwer Academic (Springer). ISBN 978-1-4020-7647-3.
- Jebara, Tony (2002). *Discriminative, generative, and imitative learning* (PhD). Massachusetts Institute of Technology. hdl:1721.1/8323 (https://hdl.handle.net/1721.1%2F8323)., (mirror (https://www.cs.columbia.edu/~jebara/papers/jebara4.pdf), mirror (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.7.513&rep=rep1&type=pdf)), published as book (above)