1.6. Data-driven approach in machine learning

*Generative modeling* is an unsupervised learning task in machine learning that involves automatically discovering and learning the regularities or patterns in input data in such a way that the model can be used to generate or output new examples that plausibly could have been drawn from the original dataset. *Generative models* try to model how data is placed throughout the space, and they are impacted by the presence of outliers. *Generative models* learn concept from a share of low level features and introduce latent variable(s) to fues various cues. They may learn object groups via shape invariance. Generative approach estimates the joint probability density function applying Baye’s rule. The reason why ‘discriminative training’ might give better results than direct use of the generative model is that it relaxes the implicit constraint (Bishop *et. al.* 2007). The main idea of generative learning is that in order to learn with understanding, a learner has to construct meaning actively. In a *generative* model, each class is learned individually and only considers the data whose labels correspond to it. The model does not focus upon inter-model discrimination and avoids considering the data as whole. A *generative* model can estimate the probability of the instance, and also the probability of a class label.

A [*discriminative model*](https://en.wikipedia.org/wiki/Discriminative_model) is a model of the [*conditional probability*](https://en.wikipedia.org/wiki/Conditional_probability) P ( Y ∣ X = x ) {\displaystyle P(Y\mid X=x)} of the target *Y*, given an observation *x*. It is also known as conditional model since it learns the boundaries between classes or labels in a dataset. *Discriminative models* are more robust to outliers. ***Discriminative*** models (also called conditional models) discriminate between different kinds of data instances (map the given unobserved variable (target) x {\displaystyle x} to a class labely {\displaystyle y} x {\displaystyle x} Y x {\displaystyle x} dependent on the observed variables (training samples)). ***Generative* models** model the **distribution** of individual classes (they try to learn the probability distribution that has generated the data) whereas ***discriminative* models** learn the (hard or soft) **boundary** between classes.

A *discriminative* model can estimate the probability that an instance belongs to a class.

*Discriminative* models try to draw boundaries in the data space, while generative models try to model how data is placed throughout the space. For example, Naïve Bayes classifier is a generative model because it uses knowledge (or assumptions) about the underlying probability distributions that generate the data being analyzed – it is capable of generating new data points. In the generative models, typically the likelihood is learned directly (and the prior and evidence approximated from the data) to calculate the posterior, whereas with *discriminative* models the posterior is learned directly. In the most basic way, generative and discriminative learning is the difference in between learning from the distribution of some data set or classes and learning from the categorical differences or boundaries between data sets or classes.

*Generative* models are often used when it is assumed that the underlying data is generated from an underlying distribution and the purpose is to find the hidden parameters of that distribution. In *discriminative* models, the best separation between the instances who fall in different classes is found. *Generative* models often outperform *discriminative* models on smaller datasets because their **generative assumptions place some structure on the model that prevents overfitting.** Discriminative models are preferred when having a lot of data. Building generative models and train them to reach a convergence is like “learning all by yourself” - it is harder but has more benefit. Though training generative models are much slower and harder than discrimitive models, they learn more and have more applications (Crossvalidated 2022).

*The generative approach* is representing the class conditional densities with a parametric model. Examples of generative classifiers are Linear Discriminant Analysis, Quadratic Discriminant Analysis. *The discriminative approach* is representing the conditional probability of a class with a semi parametric or a non-parametric model. Examples of discriminative classifiers are Logistic Regression,  
Classification trees (CART), Support Vector Machines (SVM), k-neearest neighbors. *Discriminative* models model , where is the label, and is the observed variables. A [*discriminative or conditional* model](https://en.wikipedia.org/wiki/Discriminative_model#Definition) assigns a conditional probability to one set of variables given another set of variables. Discriminative models may sometimes be trained in an unsupervised manner ([*discriminative clustering*](https://papers.nips.cc/paper/4154-discriminative-clustering-by-regularized-information-maximization)). A [*generative* model](https://en.wikipedia.org/wiki/Generative_model) assigns a joint probability distribution to all variables involved, even if we ultimately only care about a conditional or marginal distribution. Classical example of generative model is the. Naive Bayes is supervised and is usually applied to very simple data, so data complexity does not matter at all. And neither of these models was designed to generate inputs. In fact, their conditional independence assumptions make them do a poor job of it. We use "generative" and "discriminative" to quickly communicate some general properties of a probabilistic model.

Following Batch *et. al* (2017), supervised machine learning traditionally depends on access to labeled training data, a major bottleneck in developing new methods and applications. In particular, deep learning methods require tens of thousands or more labeled data points for each specific task. Collecting these labels is often prohibitively expensive, especially when specialized domain expertise is required and training data are manually labeled. Aiming to overcome this bottleneck, there is growing interest in using generative models to synthesize training data from weak supervision sources such as heuristics, knowledge bases, and weak classifiers trained directly on noisy sources. Rather than treating training labels as gold-standard inputs, such methods model training set creation as a process in order to generate training labels at scale. Batch *et. all* (2017) states that the amount of unlabeled data required to identify the true structure scales sublinearly in the number of possible dependencies (quantities decrease with the number of possible dependencies) for a broad class of models. The true class label for a data point is modeled as a latent variable that generates the observed, noisy labels. After fitting the parameters of this generative model on unlabeled data, a distribution over the latent, true labels can be inferred (Batch *et. al.* 2017). The structure of such generative models directly affects the inferred labels, and in many cases the structure is user-specified. Batch *et. al.* (2017) explains that one option is to assume that the supervision sources are conditionally independent given the latent class label (for example, in Naive Bayes model). However, statistical dependencies are common in practice, and not taking them into account leads to misjudging the accuracy of the supervision. Batch *et. al.* (2017) emphasize that we cannot rely in general on users to specify the structure of the generative model, because supervising heuristics and classifiers might be independent for some data sets but not others. We therefore seek an efficient method for automatically learning the structure of the generative model from weak supervision sources alone (Batch *et. al.* 2017). Learning the structure of generative models for weak supervision is challenging because the true class labels are latent. Although we can learn the parameters of generative models for a given structure using stochastic gradient descent and Gibbs sampling, modeling all possible dependencies does not scale as an alternative to model selection (Batch *et. al.* 2017). Batch *et. al.* (2017) and other authors propose estimators to learn the dependency structure of a generative model without using any labeled training data, but in the framework of our thesis we will use Gibbs sampling procedure which is designed to sample from an arbitrary joint distribution, in cases where it is simpler to get the conditional distribution of each element (conditional on the other elements) than it is so get the marginal distribution of the elements. Gibbs sampling iteratively samples from the conditionals of each group of variables and does this long enough in order to have samples from the joint posterior. The Gibbs sampler is applicable when we are seeking to sample from a joint distribution that is conditional on some data (i.e., a joint posterior or predictive distribution); it is also applicable when there is no (explicit) conditioning variable.