Title: Local case-control sampling

URL: https://en.wikipedia.org/wiki/Local_case-control_sampling

PageID: 46963137

Categories: Category:Logistic regression, Category:Machine learning

---

In machine learning , local case-control sampling [1] is an algorithm used to reduce the complexity of training a logistic regression classifier. The algorithm reduces the training complexity by selecting a small subsample of the original dataset for training. It assumes the availability of a (unreliable) pilot estimation of the parameters. It then performs a single pass over the entire dataset using the pilot estimation to identify the most "surprising" samples. In practice, the pilot may come from prior knowledge or training using a subsample of the dataset. The algorithm is most effective when the underlying dataset is imbalanced. It exploits the structures of conditional imbalanced datasets more efficiently than alternative methods, such as case control sampling and weighted case control sampling.

Imbalanced datasets

In classification , a dataset is a set of N data points (x i , y i) i = 1 N {\displaystyle (x_{i},y_{i})_{i=1}^{N}} , where x i ∈ R d {\displaystyle x_{i}\in \mathbb {R} ^{d}} is a feature vector, y i ∈ { 0 , 1 } {\displaystyle y_{i}\in \{0,1\}} is a label. Intuitively, a dataset is imbalanced when certain important statistical patterns are rare. The lack of observations of certain patterns does not always imply their irrelevance. For example, in medical studies of rare diseases, the small number of infected patients (cases) conveys the most valuable information for diagnosis and treatments.

Formally, an imbalanced dataset exhibits one or more of the following properties:

Marginal Imbalance . A dataset is marginally imbalanced if one class is rare compared to the other class. In other words, P (Y = 1) ≈ 0 {\displaystyle \mathbb {P} (Y=1)\approx 0} .

Conditional Imbalance . A dataset is conditionally imbalanced when it is easy to predict the correct labels in most cases. For example, if X ∈ { 0 , 1 } {\displaystyle X\in \{0,1\}} , the dataset is conditionally imbalanced if P (Y = 1 ■ X = 0) ≈ 0 {\displaystyle \mathbb {P} (Y=1\mid X=0)\approx 0} and P (Y = 1 ■ X = 1) ≈ 1 {\displaystyle \mathbb {P} (Y=1\mid X=1)\approx 1} .

Algorithm outline

In logistic regression, given the model θ = (α , β) {\displaystyle \theta =(\alpha ,\beta)} , the prediction is made according to P (Y = 1 ■ X ; θ) = p ~ θ (x) = exp ■ (α + β T x) 1 + exp ■ (α + β T x) {\displaystyle \mathbb {P} (Y=1\mid X;\theta)={\tilde {p}}_{\theta }(x)={\frac {\exp(\alpha +\beta ^{T}x)}{1+\exp(\alpha +\beta ^{T}x)}}} . The local-case control sampling algorithm assumes the availability of a pilot model θ ~ = (α ~ , β ~) {\displaystyle {\tilde {\theta }}=({\tilde {\alpha }},{\tilde {\beta }})} . Given the pilot model, the algorithm performs a single pass over the entire dataset to select the subset of samples to include in training the logistic regression model. For a sample (x , y) {\displaystyle (x,y)} , define the acceptance probability as a (x , y) = | y – p ~ θ ~ (x) | {\displaystyle a(x,y)=|y-{\tilde {p}}_{\tilde {\theta }}(x)|} . The algorithm proceeds as follows:

Generate independent z i ~ Bernoulli (a (x i , y i)) {\displaystyle z_{i}\sim {\text{Bernoulli}}(a(x_{i},y_{i}))} for i ∈ { 1 , … , N } {\displaystyle i\in \{1,\ldots ,N\}} .

Fit a logistic regression model to the subsample S = { (x i , y i) : z i = 1 } {\displaystyle S=\{(x_{i},y_{i}):z_{i}=1\}} , obtaining the unadjusted estimates θ ^ S = (α ^ S , β ^ S) {\displaystyle {\hat {\theta }}_{S}=({\hat {\alpha }}_{S},{\hat {\beta }}_{S})} .

The output model is θ ^ = (α ^ , β ^) {\displaystyle {\hat {\theta }}=({\hat {\alpha }},{\hat {\beta }})} , where α ^ ← α ^ S + α ~ {\displaystyle {\hat {\alpha }}\leftarrow {\hat {\alpha }}_{S}+{\tilde {\alpha }}} and β ^ ← β ^ S + β ~ {\displaystyle {\hat {\beta }}\leftarrow {\hat {\beta }}_{S}+{\tilde {\beta }}} .

The algorithm can be understood as selecting samples that surprises the pilot model. Intuitively these samples are closer to the decision boundary of the classifier and is thus more informative.

Obtaining the pilot model

In practice, for cases where a pilot model is naturally available, the algorithm can be applied directly to reduce the complexity of training. In cases where a natural pilot is nonexistent, an estimate using a subsample selected through another sampling technique can be used instead. In the original paper describing the algorithm, the authors propose to use weighted case-control sampling with half the assigned sampling budget. For example, if the objective is to use a subsample with size N = 1000 {\displaystyle N=1000} , first estimate a model θ ~ {\displaystyle {\tilde {\theta }}} using N h = 500 {\displaystyle N_{h}=500} samples from weighted case control sampling, then collect another N h = 500 {\displaystyle N_{h}=500} samples using local case-control sampling.

Larger or smaller sample size

It is possible to control the sample size by multiplying the acceptance probability with a constant c {\displaystyle c} . For a larger sample size, pick c > 1 {\displaystyle c>1} and adjust the acceptance probability to min (c a (x i , y i) , 1) {\displaystyle \min(ca(x_{i},y_{i}),1)} . For a smaller sample size, the same strategy applies. In cases where the number of samples desired is precise, a convenient alternative method is to uniformly downsample from a larger subsample selected by local case-control sampling.

Properties

The algorithm has the following properties. When the pilot is consistent , the estimates using the samples from local case-control sampling is consistent even under model misspecification . If the model is correct then the algorithm has exactly twice the asymptotic variance of logistic regression on the full data set. For a larger sample size with c > 1 {\displaystyle c>1} , the factor 2 is improved to 1 + 1 c {\displaystyle 1+{\frac {1}{c}}} .

References