Intensionality of Adjectives

Mark Kowarsky MARKAK@STANFORD.EDU
Neha Nayak NAYAKNE@STANFORD.EDU

1. Introduction

Given the two sentences "this fake cat has a fluffy tail" and "this tabby cat has a fluffy tail", only the second allows one to conclude that "there exists a cat that has a fluffy tail". As is obvious to a native English speaker, the adjective (A) "fake" somehow modifies the noun (N) "cat" such that we conclude that it no longer belongs to the set of N (cats) by virtue of being A (fake). Linguists categorize adjectives with this effect as being *intensional*.

A noun N described by an intensional adjective may not belong to "the set of Ns". An intensional adjective is either privative - it precludes the N it modifies from belonging to the set of Ns. Or plain non-subsective - it conveys uncertainty about whether the N it modifies belongs to the set of Ns.

Privative adjectives :	*former, mock, false*
Plain non-subsective adjectives :	*alleged, ostensible*

Table 1. Intensional adjective subclasses

Distinguishing between intensional and extensional (non-intensional) adjectives is an important part of understanding textual entailment, essential to the domain of Natural Language Understanding (Angeli & Manning, 2013). The following sentences illustrate how this problem is relevant to logical inference. Given the information *"The young president has been travelling for eight weeks"*, we can correctly conclude that *"The president has been travelling for eight weeks"* However, if it is known that *"The former president has been travelling for eight weeks"*, it would be incorrect to conclude that *"The president has been travelling for eight weeks"*, since the former president is not the president.

Related work

In Boleda et al. (2012), the distributional representations of various intensional and extensional adjectives were examined. Different composition functions were used to model adjectival modification, and the predicted properties of the compositional representations were evaluated. Using the cosine similarity measure, the highest degree of similarity between the observed vectors of the adjective and adjective-noun pair was for intersective adjectives, the lowest for intensional adjectives. Intensional adjective-noun vectors were most similar to those of the unmodified noun. This was attributed to the reduced potential contexts for intensional adjectives, as compared to intersective ones.

Boleda et al. (2013) conducted a similar experiment to their earlier work, using an enhanced list of intensional adjectives, as well as a more varied set of extensional adjectives. Both concluded that for the purposes of modelling adjectival modification, intensional and extensional adjectives were both modelled equivalently well with existing composition functions.

This work

This final project for CS224N and CS229, will attempt to classify whether an unseen adjective is intensional based on a linguistic model and to expand the list of intensional adjectives known in the literature. It increases the number of known intensional adjectives by 120%. It does this using adjective-noun co-occurrences (section 4), co-occurrences of adverbs modifying the adjectives (section 5) and other grammatical and contextual information (section 6) and the machinery of support vector machines.

2. Data

The three data sets we extracted for this report all derive from dumps of English Wikipedia articles. One database which had undergone tokenization, part-of-speech tagging and sentence splitting (carried out by the NLP group at Stanford) was used for adjective-noun co-occurrences. In total, 19GB of uncompressed text was used, covering 2019 different types of adjective (188 075 unique), 26 728 types of noun (172 090 unique) and total co-occurrences of 16 996 971. The other database had dependency and token contexts annotated. This allowed adverbs (1428 unique) modifying adjectives (1302 unique) to be extracted for a total of 644 052 co-occurrences. Finally, 15 077 adjective-noun pairs were extracted with 371 582 different contexts, for a total of 16 027 099 co-occurrences.

The list of known intensional adjectives (see Appendix A) was curated by reading the literature (Boleda et al., 2013), and expanded by adding synonyms as well as false positives found during early testing of the classifier[1].

3. Methods

Labelled data

The problem we are trying to solve is a supervised learning classification problem. The fact that we had only 30 known intensional adjectives had the following repercussions:
Incorrect example labels We assumed that all adjectives not known to be intensional were extensional. This simplifying assumption was obviously incorrect, and would have affected our classification results. This was confirmed by examining the false positives produced by our initial classifiers. We attempted to address this problem by relabelling training data and retraining on it.
Train-Test division We did not have enough positive examples to make a reasonable Train-Dev-Test split. We split the known intensional adjectives into a set of 10 for testing and 20 for training, and used up to 1000 and 2000 additional extensional adjectives for testing and training, respectively. All the learning algorithms applied to the training data were used with stratified k-fold cross-validation, which maintains the ratio of positive to negative examples in the training and test sets of each fold.

Classifier

A linear support vector machine (SVM) was used (Pedregosa et al., 2011; Fan et al., 2008) to classify the tokens as being either intensional or extensional. Before running the SVM, the input data was scaled to have zero mean and unit variance, which had the benefit of both increasing our precision and recall as well as speeding up the algorithm. The penalty for errors for each class was set to be inversely proportional to their frequency in the training data, to account for the different ratios of the classes.

Features

Co-occurrences are integer frequencies, which on their own do not adequately represent the significance of a particular co-occurrence. We wish to use the co-occurrence data to discriminate between the different adjective classes, and so we applied some standard transformations to the data before classification, with the objective of filtering out less meaningful co-occurrences. The following transformations were used:

$$\mathrm{PMI}(A,N) = \log(\frac{\Pr(A \cap N)}{\Pr(A) \cdot \Pr(N)})$$

$$\mathrm{LMI}(A,N) = \Pr(A \cap N) \cdot log(\frac{\Pr(A \cap N)}{\Pr(A) \cdot \Pr(N)})$$

$$\mathrm{PMI}^2(A,N) = \log(\frac{\Pr(A \cap N)^2}{\Pr(A) \cdot \Pr(N)})$$

PMI did not perform better than the raw frequencies in any of our classifiers, which was expected because of its tendency to weight very rare events favourably. Both LMI and PMI^2 should not have this effect, as they are directly related to the actual frequencies of the pairs.

Feature selection

To minimize the effect of over-fitting with too many features and to provide a meaningful subset of features for understanding how intensional adjectives are use, the following methods.

Based on association measures We used a thresholding scheme to select nouns whose top-x co-occurrences include some minimum number of intensional adjectives. This gave a list of nouns which appear to be more characteristic of intensionality, such as:

infringement	{alleged, past, potential, apparent}
perpetrator	{alleged, likely, potential, possible}
cure	{possible, potential, alleged, apparent}

Table 2. Nouns associated with multiple intensional adjectives

We trained a classifier on a subset of the features corresponding to the nouns selected by this criterion. This classifier was superseded by one in which feature selection was carried out using the Differential Expression (DE) analysis.

Differential Expression Another approach to categorising and building models that predict the class

[1] Examples found by early classifiers were strictly restricted to the training set in the classifiers used for evaluation.

of adjective has been to appropriate techniques used in bioinformatics to find nouns (genes) that are called "differentially expressed" between different adjectives (tissues/samples)(Robinson et al., 2010). The model used assumes a negative binomial model for counts which whilst not being motivated by natural language considerations is a distribution that can resemble Zipf's law so is plausibly appropriate to apply.

Briefly, it works by performing statistical tests between different classes for a given feature, taking into account the variance both within and between classes to measure the likelihood that a feature is expressed higher or lower (p-value<0.05, after Benjamini-Hochberg corrections for multiple testing(Benjamini & Hochberg, 1995)). The other method is to sort features by their "log fold change". That is, after suitable normalizations, how much are they expressed in one class compared to another. It was observed that picking features using the log fold-change method performed best.

In an attempt to improve the list of features again, an automatic recursive feature elimination with cross-validation algorithm was also employed to minimize the set, using the

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

as the metric of the quality of the fit.

apartheid {anti-,former}
contradiction {apparent, seeming}
lack {alleged, apparent, former, likely, past, possible, potential, probable, seeming, virtual}

Table 3. Some nouns associated with the intensional class - selected by DE

4. Identifying nouns indicative of intensionality

Our baseline implements a simple hypothesis: that intensional and extensional adjectives differ in the nouns they can modify. Using adjective-noun bigrams, we constructed a vector space of adjectives, using their co-occurrence frequencies (above a threshold) with nouns as features.

The best results were achieved by using nouns (features) chosen by log fold-change as determined by differential expression and counts modified by PMI^2. Using the recursive feature elimination with cross-validation increased the number of false positives and false negatives, so was not used.

We updated the training set by incorporating false positive adjectives (see Table 4.2) that turned out to be mislabelled to attempt to better train the learning algorithm.

4.1. Results

We trained classifiers on sets different ratios of extensional and intensional adjectives ($m_{ext} : m_{int}$), using randomly selected subsets of the training set. Using a different number of features selected by DE, and different $m_{ext} : m_{int}$ ratios, we produced learning curves of the average precision, recall and F_1 scores over the stratified test folds. The solid lines in the plots below is for the original training set, the dashed line when we relabelled false positives that were in fact intensional adjectives.

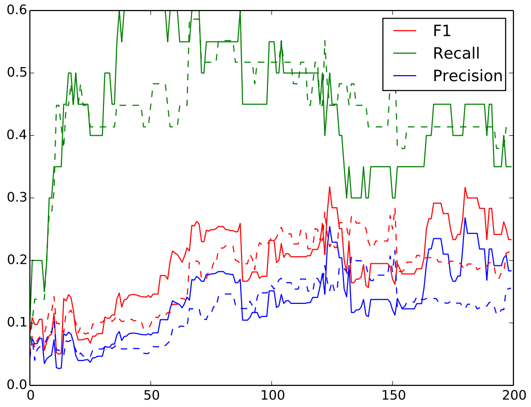

Figure 1. Learning curve for 100:1 ratio of intensional:extensional adjectives

The confusion matrices in Table 12 show more detailed results of the classifiers. Each cell contains the average of the corresponding cells in the confusion matrix of the folds of k-fold cross validation. The matrices labelled *With 'Bootstrapping'* correspond to the classifiers in which we manually flipped the labels on the false positives that were actually intensional.

The first classifier, with an $m_{ext} : m_{int}$ ratio of 100:1, represents a model of the 'real-life' distribution of adjectives between the classes. In this situation, we aimed to increase recall while maintaining high precision - corresponding to the task of gathering as many unknown intensional adjectives as possible. We were not able to beat the majority baseline of 99% in this scenario.

The third classifier, using an $m_{ext} : m_{int}$ ratio of 1:1, represents an attempt to characterise fundamental

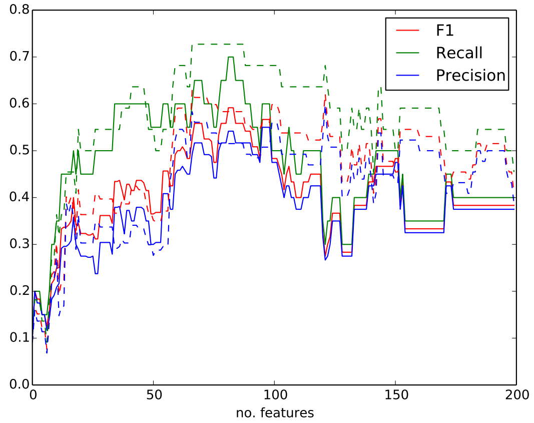

Figure 2. Learning curve for 10:1 ratio of intensional:extensional adjectives

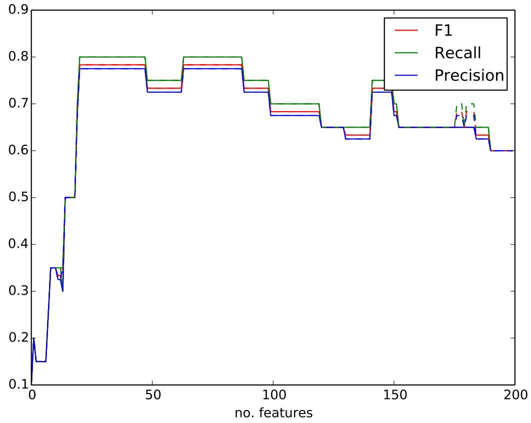

Figure 3. Learning curve for 1:1 ratio of intensional:extensional adjectives

differences between intensional and extensional adjectives. With an accuracy of 85%, this classifier was able to beat the majority baseline of 50%. The constant values post-'bootstrapping' occurred because the training examples whose labels were flipped did not occur in the training set for this classifier.
The second classifier has an $m_{ext} : m_{int}$ ratio of 10:1, and represents an intermediate case.

Initial		Predicted Class	
		Ext	Int
True Class	Ext	96.85	3.10
	Int	0.4	0.6
With 'Bootstrapping'		Predicted Class	
		Ext	Int
True Class	Ext	65.28	3.34
	Int	0.45	0.55

Table 4. Confusion matrix - Best classifier for 100:1. 82 nouns selected by DE

Initial		Predicted Class	
		Ext	Int
True Class	Ext	9.45	0.55
	Int	0.35	0.65
With 'Bootstrapping'		Predicted Class	
		Ext	Int
True Class	Ext	8.41	0.59
	Int	0.27	0.73

Table 5. Confusion matrix - Best classifier for 10:1. 67 nouns selected by DE

4.2. Analysis

False negatives Polysemous adjectives with extensional meanings but labelled 'intensional' in our data were frequently mislabelled. For example, this occurred with 'theoretical' (which often modifies 'physics') and 'artificial' (which often modifies 'intelligence' and 'insemination'). Adjectives such as 'likely' and 'probable' - with exactly one meaning, which is intensional - tended to fare better, and were classified as extensional less frequently. Our classifiers also misclassified 'phony' and 'erstwhile' as extensional, which may be due to the lack of data, stemming from their low frequencies.

False Positives Manually sifting through the false positives revealed many intensional adjectives that we had not considered in our seed list. Two conclusions can be drawn from this. First, the accuracy of our system may be higher than the confusion matrices show. Second, our system recovers intensional adjectives at a rate higher than chance.

Insights from error analysis prompted us to return to our original list of intensional adjectives to find characteristics that might make some of them problematic, which would then guide the development of features for subsequent classifiers.

Polysemous words Certain adjectives in our seed set which were labelled as intensional were actually only intensional when used in a certain sense. For example, assumed culprit may not actually be a culprit, but an assumed name is, in fact, a name. This led us to the inference that intensionality is the characteristic of a particular instance of adjectival modification, and

Initial		Predicted Class	
		Ext	Int
True Class	Ext	0.9	0.1
	Int	0.2	0.8
With ‘Bootstrapping’		Predicted Class	
		Ext	Int
True Class	Ext	0.9	0.1
	Int	0.2	0.8

Table 6. Confusion matrix - Best classifier for 1:1. 64 nouns selected by DE

historic, faulty, uncertain, plausible, erroneous, unlikely, suspicious, unsuccessful, illegitimate, simulated

Table 7. False positives

informed the classifiers in 6.

Idioms and common collocations While the word ‘potential’ in the context of ‘potential solution’ is intensional, ‘potential’ occurs most commonly in the corpus as part of the collocation ‘potential energy’, which is, in fact, subsective. Also, the idioms such as ‘false alarm’ occur with high frequencies and are not examples of intensional modification. This is further justification for instance-based classification.

Antonyms Many antonyms of privative adjectives were wrongly classified as intensional. These are adjectives that affirm the membership of a noun to its class, and appear in similar context to privative adjectives. It was not expected that instance-based classification could correctly classify these instances, since the ambiguity was inherent in the contexts. However, it was worth noting that these adjectives might cause persistent errors. Some examples include:

true, complete, obvious, factual

Table 8. Antonyms of privative adjectives, mislabelled

5. Identifying adverbs indicative of intensionality

We carried out classifications parallel to those in 4, using the co-occurrence of adverbs instead of nouns. We took advantage of the dependency-parsed Wikipedia corpus, which allowed us to consider long-range dependencies by considering arcs between adverbs and adjectives with the label *advmod*.

Unlike the nouns, this did not produce useful results. Even when training on all of the adverbs, the training set would often have an accuracy below the baseline and almost all test sets did not recover any of the intensional adjectives. The lack of results motivated us to try for other features relating to the context in which the adjective appears, detailed in the next section.

6. Instance-based classification

Method

From the errors in 4, we concluded that although certain adjectives are usually intensional, most of them are not intensional in all contexts, for example: *false alarm, assumed identity.* We modified our hypothesis, claiming that intensionality is a characteristic of particular instances of adjectival modification. In the instance-based classifier, each example corresponds to an adjective-noun pair.

One advantage of this model was that we were no longer restricted to a maximum of 30 positive examples, since a positive example could be constructed out of any occurrence of an intensional adjective. We also hoped to use the model to identify instances of ’intensional modification’ that do not necessarily contain an occurrence of a known intensional adjective. This would be useful for the overarching task of identifying problematic adjectives for logical inference.

For each adjective in the training and test sets, we selected five distinct nouns most frequently modified by it (or as many as were available, if fewer). All adjective-noun pairs in which the adjective was usually intensional were labelled as intensional, although this contradicts the idea behind instance-based classification. We hoped to find reliable false positives that would allow bootstrapping, which would address this problem.

We hypothesised that since a noun N modified by an intensional adjective is no longer ‘an N’, the contexts in which the modified and unmodified noun occur would be distinct. For every pair of modified noun and context (A-N, y) that co-occurred in the corpus, we calculated

$$f(A\text{-}N, y) = \frac{\Pr(y \mid A\text{-}N)}{\Pr(y \mid \text{unmodified-}N)}.$$

Using add-one smoothing for the probabilities of the unmodified contexts (since is it possible that they might be zero), this can be expressed as

$$f(A\text{-}N, y) = \frac{c(y \cap A\text{-}N) \cdot (c(\text{unmodified-}N) + |Y|)}{c(A\text{-}N) \cdot c(\text{unmodified-}N \cap y)},$$

where c is a counting function. We also calculated the following score, which was constrained between $[-1, 1]$

$$g(A\text{-}N, y) = \frac{c(y \cap A\text{-}N) - c(y \cap \text{unmodified-}N)}{c(y)}$$

These functions were chosen to differentiate between contexts that occurred ‘only with A-modified Ns’ and that occurred ‘only with unmodified Ns’. In Boleda et al. (2012), the cosine measure showed high similarity between the contexts of intensionally modified Ns and unmodified Ns, since the contexts of the intensional nouns were a subset. Our measure does not have this property. We also do not expect the contexts of intensionally modified nouns to be a subset of those of the corresponding unmodified nouns, since we are using fixed window-based contexts as features, rather than a bag-of-words representation.

The experiments were conducted using token-based contexts, with a window of up to two tokens on either side of the target.

To reduce the number of features used by the classifier, we considered only the top 5000 context features by frequency. Since these were not always informative, we also considered the top 5000 context features after weighting with PMI^2. We also modified the experiment to include a threshold for the number of times a context must co-occur with a noun or adjective-noun instance to be considered.

Results

The instance-based classifiers performed very poorly, achieving accuracy below the majority baseline for all ratios of $m_{extensional} : m_{intensional}$. Without further investigation, this could be due to the assumption made about our training data, that if the adjective is marked as intensional then the adjective-noun pair is as well. A more nuanced training set may improve the results and reduce the large number of false positives and negatives we receive. Results for a representative subset of classifiers are shown in Table 9.

	F1	R	P
f, top 5k contexts	4.05%	29.5%	2.21%
g, top 5k contexts	7.98 %	23.87%	4.89%
f, strict threshold	10.3%	20.45%	7.14%
g, strict threshold	6.34%	18.18%	3.90%
f, PMI^2 contexts	0.20%	9.09%	0.10%
g, PMI^2 contexts	0.20%	9.09%	0.10%

Table 9. Results for instance-based classification.

Analysis

Zipf’s Law The frequency distribution of the contexts observed was similar to that predicted by Zipf’s law. The patterns that occurred with high frequencies were relatively uninformative, for example,

```
the__*  969989
*__of   767112
a__*    657886
*__,    438557
*__.    433165
```

Frequencies then dropped off sharply, and a large subset of the distinct patterns occurred only once, or co-occurred with exactly one Adjective-Noun pair. This issue could be addressed by introducing generalisations in the patterns, for example, replacing words with their direct or indirect hypernyms from WordNet. For example, replacing the contexts ‘American__*’, ‘British__*’ and other nationalities with ‘<Nationality >__*’, which would reduce sparsity without sacrificing meaning.

Selection of adjective-noun bigrams It is possible that more judicious selection of adjective-noun bigrams would have been more appropriate. Selecting the five nouns that co-occur most frequently with each adjective may have resulted in the inclusion of idiomatic expressions, or collocations that do not strictly correspond to adjectival modification, for example, *Big Brother*.

Feature selection and transformations In the noun co-occurrence experiments, we used transformations to ensure that more meaningful co-occurrences were given a higher weight. However, such transformations in the instance-based classification data would render the function f meaningless, as it is defined for probabilities. PMI^2 did not help to extract meaningful co-occurrences, resulting in classifiers that indiscriminately labelled instances as extensional. A more effective measure by which to weight pair-context co-occurrences would contribute to more effective feature selection.

7. Evaluation

As only the adjective-noun co-occurrence classifier performed reasonably and features from the other classifiers were uninformative, we chose to use only the adjective-noun co-occurrence data in our final training set. The linear SVM was trained against all the training data using the appropriate number of features as found maximized the number of correct predictions in the training set.

Initial		Predicted Class	
		Ext	Int
True Class	Ext	921	79
	Int	9	1

Table 10. Confusion matrix - Final evaluation for 100:1. 82 nouns selected by DE

Initial		Predicted Class	
		Ext	Int
True Class	Ext	87	13
	Int	9	1

Table 11. Confusion matrix - Final evaluation for 10:1. 67 nouns selected by DE

Initial		Predicted Class	
		Ext	Int
True Class	Ext	9	1
	Int	7	3

Table 12. Confusion matrix - Final evaluation for 1:1. 64 nouns selected by DE

These confusion matrices show that we often only identify 10% of the intensional adjectives in the test set. Increasing the number of features we trained upon increases this percentage for the lower extensional:intensional ratio samples, indicating that it may just be doing an improved job at classifying those adjectives as being intensional against the *particular* extensional adjectives, rather than the class of all of them.

Further work could divide the intensional adjectives into a few subclasses based on their semantic properties (for example, separating privative and plain non-subsective adjectives) and try to train on each of those. This would hopefully better capture the variety of ways different classes of adjectives are used. Although the instance based features did not work well here, finding a good choice and training with appropriately labelled training data, data that does not assume every instance of an intensional adjective is used in an intensional manner, still feels like it may be useful.

Even though we were not highly successful in classifying unseen adjectives, we did manage to increase the number of known intensional adjectives by 120%.

Acknowledgments

We would like to thank Gabor Angeli for the original project idea of detecting intensionality and for the numerous discussions we had with him during the course of this project. He was always good to bounce ideas off of and also provided us with some of the scripts needed to extract the data from the corpora.

A. Intensional adjectives

From literature: possible, potential, apparent, likely, theoretical, alleged, hypothetical, probable, presumed, putative, former, future, past, false, artificial, impossible, mock, fake, counterfeit, fictitious, ostensible
From synonyms: ex-, phony, virtual, vice, adjunct, unlikely, unnecessary
From false positives: anti-, assumed, mistaken, erstwhile, fictional, seeming, deputy, associate, historic, faulty, uncertain, erroneous, plausible, suspicious, unsuccessful, illegitimate, simulated

References

Angeli, Gabor and Manning, Christopher D. Philosophers are mortal: Inferring the truth of unseen facts. *CoNLL-2013*, pp. 133, 2013.

Benjamini, Yoav and Hochberg, Yosef. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300, 1995.

Boleda, Gemma, Vecchi, Eva Maria, Cornudella, Miquel, and McNally, Louise. First-order vs. higher-order modification in distributional semantics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1223–1233. Association for Computational Linguistics, 2012.

Boleda, Gemma, Baroni, Marco, and McNally, Louise. Intensionality was only alleged: On adjective-noun composition in distributional semantics. In *Proceedings of IWCS*, pp. 35–46, 2013.

Fan, Rong-En, Chang, Kai-Wei, Hsieh, Cho-Jui, Wang, Xiang-Rui, and Lin, Chih-Jen. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Robinson, Mark D, McCarthy, Davis J, and Smyth, Gordon K. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.