به نام خدا

۱ classification

۱.۱ ارزیابی

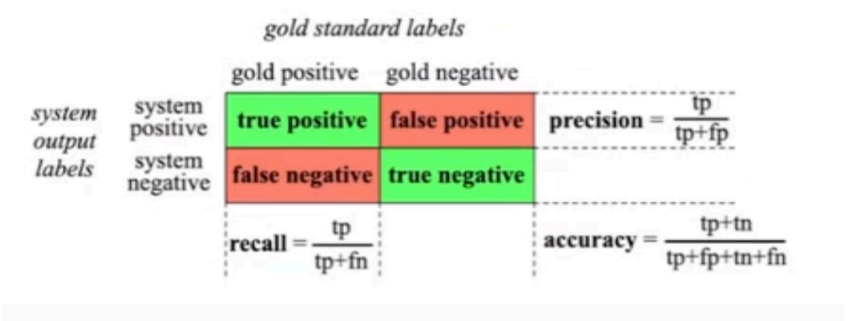
برای ارزیابی از Cross-Validation استفاده میکنیم به این صورت که داده ها را به k بخش قسمت کرده در هر که در هر بخش یک قسمت را به داده های test و مابقی را به داده های train بخش بندی میکنیم در نهایت مدلی که بیشترین میانگین و کمترین واریانس را دارد به عنوان بهترین مدل در نظر میگیریم

*معیار های ارزیابی عبارتند از :

$$\text{Microf}\,۱ = \sum_c \alpha_i f۱_c,\ \text{Macro}\ f۱ = \frac{۱}{|c|}\sum_c f۱_c$$

**α_i بر اساس تعداد داده ها

*سایر معیار های ارزیابی در شکل ۱ آمده اند.



شکل۱:معیارهای ارزیابی

۲.۱ $GenerativeClassifiers - NaveBayes$

$$\text{Buyes: Naive}\ \hat{y} = \underset{y\in Y}{\text{argmax}} P(y \mid x) = \underset{y\in Y}{\text{argmax}} \frac{P(x|y)P(y)}{P(n)\rightsquigarrow \text{ignore}}$$

اگر حالت unigram در نظر بگیریم $P(n \mid y) = \prod_{i=۱}^{n} P(w_{v_i} \mid y)\ P(y)$

و اگر تعداد کلمات $|V|$ باشد برای یادگیری آن $۲|V| + ۱$ پارامتر لازم داریم

۳.۱ $DiscriminativeClassifier - logisticregression$

$$w.x = \sum_{i=۱} w_i x_i\ \rho(y = ۱) = \sigma(w \cdot x + b) = \hat{y}$$

$$\rho(y = ۰) = ۱ - \rho(y = ۱) = ۱ - \hat{y}$$

$$p\left(y(x) = \hat{y}^y \times (۱ - \hat{y})^{۱-y}\right) \rightarrow -\log p(y \mid x) = -(y \log \hat{y} + (۱ - y)\log(۱ - \hat{y})) = L_{CE}\left(y, \hat{y}'\right). \rightarrow \text{loss crossentropy}$$

برای جلو گیری از overfitting:

$$\hat{\theta} = \underset{\theta}{\text{argmax}} \sum_{i=۱}^{m} \log p\left(y^i \mid x^i\right) - \underbrace{\alpha R(\theta)}_{\text{regularization}} \quad R(\theta) = \begin{cases} \|\theta\|_۲^۲ = \sum \theta_i^۲ \\ \|\theta\|_۱ = \sum |\theta_i| \end{cases}$$

۲ تحلیل sentiment

**تحلیل احساسات جنبه های زیادی دارد و میتواند در سطح doc ، token ، ... بررسی شود

* بحث sentiment و aspect خیلی باهم آمیخته هستند ولی معمولا این دوتا رو باهم برچسب زده نداریم؛

به عبارت دیگر مثلا ما یک کلمه را نداریم که از یک منظر برچسب مثبت و از یک منظر برچسب منفی داشته باشه

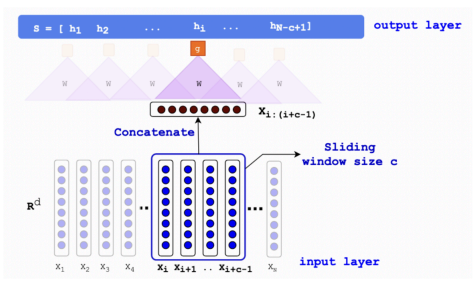
*راه های شناسایی تشخیص مثبت یا منفی بودن جمله عبارتند از :

۱.تعداد کلملات مثبت و منفی را می شماریم اگر تعداد مثبت ها بیشتر بود جمله مثبت و اگر تعداد منفی ها بیشتر بود جمله منفی .

مشکل: احتمال دارد تعداد مثبت و منفی یکی باشد در نتیجه جمله خنثی در نظر گرفته میشود

۲. استفاده از Bayes Naive که فقط کلمات مثبت منفی را در نظر بگیریم یا کل کلمات رو در نظر بگیریم

۳.استفاده از Regression Logistic یا استفاده از network neural ها



شکل۲ :استفاده از cnn برای حل مسئله طبقه بندی

**اگر Multiclass بود از Softmax در تابع فعال سازی لایه اخر استفاده مکینم اگر multilabel ، Multiclass بود از sigmoid .