Точность, полнота, F-мера - оценка классификатора

Тестовая выборка

Основой проверки является тестовая выборка, в которой проставлено соответствие между документами и их классами.

В зависимости от ваших конкретных условий получение подобной выборки может быть затруднено, так как зачастую ее составляют люди. Но иногда ее можно получить без большого объема ручной работы.

Когда у вас появилась тестовая выборка достаточно натравить классификатор на документы и соотнести его решение с заведомо известным правильным решением. Но для того чтобы принимать решение хуже или лучше справляется с работой новая версия алгоритма нам необходима численная метрика его качества.

Численная оценка качества алгоритма

Accuracy

В простейшем случае такой метрикой может быть доля документов, по которым классификатор принял правильное решение.

$$\text{Accuracy} = \frac{P}{N}$$

где, P – количество документов, по которым классификатор принял правильное решение, а N – размер обучающей выборки. Очевидное решение, на котором для начала можно остановиться.

Тем не менее, у этой метрики есть одна особенность, которую необходимо учитывать. Она присваивает всем документам одинаковый вес, что может быть не корректно в случае, если распределение документов в обучающей выборке сильно смещено в сторону какого-то одного или нескольких классов. В этом случае у классификатора есть больше информации по этим классам и соответственно в рамках этих классов он будет принимать более адекватные решения. На практике это приводит к тому, что вы имеете **Accuracy**, скажем, 80%, но при этом в рамках какого-то конкретного класса классификатор работает из рук вон плохо не определяя правильно даже треть документов.

Один выход из этой ситуации заключается в том, чтобы обучать классификатор на специально подготовленном, сбалансированном корпусе документов. Минус этого решения в том, что вы отбираете у классификатора информацию об относительной частоте

документов. Эта информация при прочих равных может оказаться очень кстати для принятия правильного решения.

Другой выход заключается в изменении подхода к формальной оценке качества.

Точность и полнота

Точность (**Precision**) и полнота (**Recall**) являются метриками, которые используются при оценке большей части алгоритмов извлечения информации. Иногда они используются сами по себе, иногда в качестве базиса для производных метрик, таких как F-мера или R-Precision. Суть точности и полноты очень проста.

Точность системы в пределах класса – это доля документов, действительно принадлежащих данному классу относительно всех документов, которые система отнесла к этому классу.

Полнота системы – это доля найденных классификатором документов, принадлежащих классу относительно всех документов этого класса в тестовой выборке.

Эти значения легко рассчитать на основании таблицы контингентности, которая составляется для каждого класса отдельно.

Категория i		Экспертная оценка	
		Положительна	Отрицательная
Оценка системы	Положительная	TP	FP
	Отрицательная	FN	TN

В таблице содержится информация сколько раз система приняла верное и сколько раз неверное решение по документам заданного класса. А именно:

- TP — истинно-положительное решение;
- TN — истинно-отрицательное решение;
- FP — ложноположительное решение;
- FN — ложноотрицательное решение.

Тогда, точность и полнота определяются следующим образом:

$$\textbf{Precision}=\frac{TP}{TP+FP}$$

$$\textbf{Recall}=\frac{TP}{TP+FN}$$

Рассмотрим пример.

Допустим, у вас есть тестовая выборка, в которой 10 сообщений, из них 4 – спам.

Обработав все сообщения классификатор пометил 2 сообщения как спам, причем одно действительно является спамом, а второе было помечено в тестовой выборке как нормальное. Мы имеем

одно истинно-положительное решение, три ложноотрицательных и одно ложноположительное. Тогда для класса “спам” точность классификатора составляет $\frac{1}{2}$ (50% положительных решений правильные), а полнота $\frac{1}{4}$ (классификатор нашел 25% всех спам-сообщений).

Confusion Matrix

На практике значения точности и полноты гораздо более удобней рассчитывать с использованием матрицы неточностей (confusion matrix). В случае если количество классов относительно невелико (не более 100-150 классов), этот подход позволяет довольно наглядно представить результаты работы классификатора.

Матрица неточностей – это матрица размера N на N, где N — это количество классов.

Столбцы этой матрицы резервируются за экспертными решениями, а строки за решениями классификатора.

Когда мы классифицируем документ из тестовой выборки мы инкрементируем число, стоящее на пересечении строки класса, который вернул классификатор и столбца класса, к которому действительно относится документ.

	0.91	0.96	0.94	0.75	1.00	0.83	0.85	0.97	1.00	0.86	1.00	0.79	1.00	0.75	1.00	1.00	0.96	0.90	0.81	0.89	0.94	0.98	0.86	0.89	0.94	0.92	0.96
0.80		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
0.95	1	94	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
1.00	2	0	32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.29	3	0	0	6	0	0	3	2	0	1	0	0	0	0	0	0	1	1	0	0	1	0	1	3	0	2	0
1.00	4	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.50	5	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	2	0	1	1
0.92	6	1	0	0	0	0	152	0	0	1	0	0	0	0	0	0	0	1	4	2	3	0	0	0	0	2	0
0.97	7	1	0	1	0	0	0	256	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	2	0
0.33	8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0
0.97	9	0	0	0	0	0	0	0	0	69	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
0.82	10	0	0	0	0	0	2	0	0	0	18	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
0.87	11	0	0	0	0	0	0	0	0	0	0	34	0	4	0	0	0	0	0	0	0	0	0	1	0	0	0
1.00	12	0	0	0	0	0	0	0	0	0	0	0	37	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.57	13	0	0	0	0	0	0	0	0	0	0	9	0	12	0	0	0	0	0	0	0	0	0	0	0	0	0
0.63	14	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	3	0	0	0	0	0	0	0	0	0
0.50	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	1	1	0	0	0	0	0	0
0.77	16	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	47	0	1	3	4	0	0	2	0	1	0
0.87	17	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	69	1	2	5	0	0	0	0	0	0
0.97	18	0	0	0	0	1	4	0	0	1	0	0	0	0	0	0	0	0	197	1	0	0	0	0	0	0	0
0.78	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	35	183	13	0	0	2	0	1	0
0.97	20	0	0	0	0	0	10	3	0	1	0	0	0	0	0	0	0	0	0	4	702	0	0	0	0	6	0
0.93	21	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	56	0	2	0	0	0
0.29	22	0	0	1	0	0	2	0	0	6	0	0	0	0	0	0	0	0	1	1	1	0	6	2	0	1	0
0.91	23	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	3	6	0	0	115	0	0	0
1.00	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0
0.93	25	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2	4	5	0	0	0	1	196	0
0.98	26	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	78

Матрица неточностей (26 классов, результирующая точность – 0.8, результирующая полнота – 0.91)

Как видно из примера, большинство документов классификатор определяет верно. Диагональные элементы матрицы явно выражены. Тем не менее в рамках некоторых классов (3, 5, 8, 22) классификатор показывает низкую точность.

Имея такую матрицу точность и полнота для каждого класса рассчитывается очень просто.

Точность равняется отношению соответствующего диагонального элемента матрицы и суммы всей строки класса.

Полнота – отношению диагонального элемента матрицы и суммы всего столбца класса.

Формально:

$$\mathbf{Precision}_c = \frac{A_{c,c}}{\sum_{i=1}^{n} A_{c,i}}$$

$$\mathbf{Recall}_c = \frac{A_{c,c}}{\sum_{i=1}^{n} A_{i,c}}$$

Результирующая точность классификатора рассчитывается как арифметическое среднее его точности по всем классам.

То же самое с полнотой.

Технически этот подход называется **macro-averaging**.

F-мера

Понятно, что чем выше точность и полнота, тем лучше. Но в реальной жизни максимальная точность и полнота не достижимы одновременно и приходится искать некий баланс. Поэтому, хотелось бы иметь некую метрику, которая объединяла бы в себе информацию о точности и полноте нашего алгоритма. В этом случае нам будет проще принимать решение о том какую реализацию запускать в **production** (у кого больше тот и лучше). Именно такой метрикой является F-мера.

F-мера представляет собой гармоническое среднее между точностью и полнотой. Она стремится к нулю, если точность или полнота стремится к нулю.

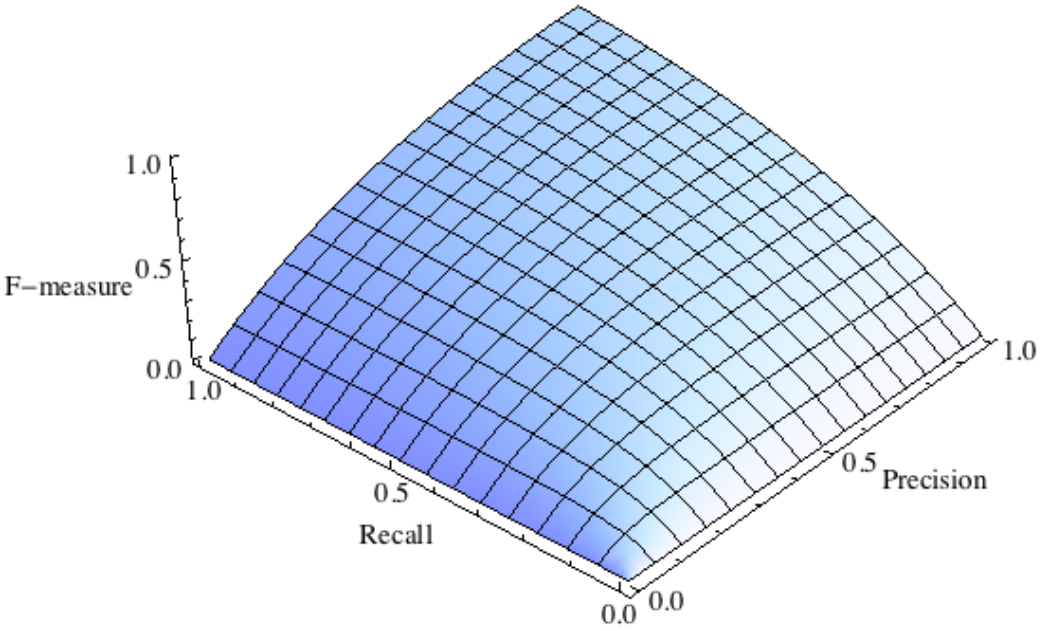
$$\mathrm{F} = 2\frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

Данная формула придает одинаковый вес точности и полноте, поэтому F-мера будет падать одинаково при уменьшении и точности и полноты.

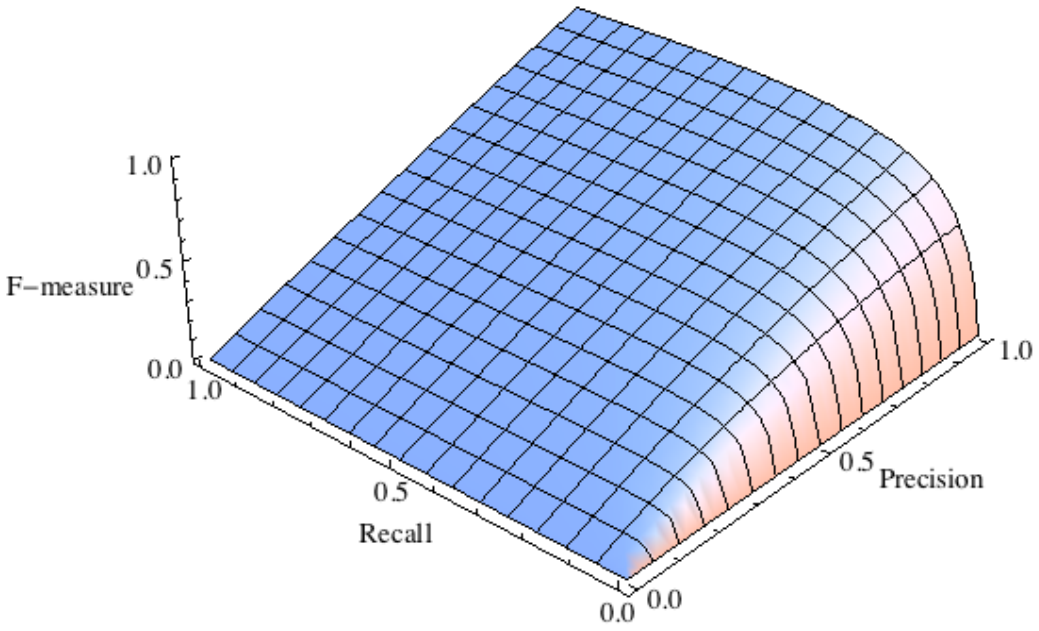
Возможно рассчитать F-меру придав различный вес точности и полноте, если вы осознанно отдаете приоритет одной из этих метрик при разработке алгоритма.

$$F = (\beta^2 + 1)\frac{Precision \times Recall}{\beta^2 Precision + Recall}$$

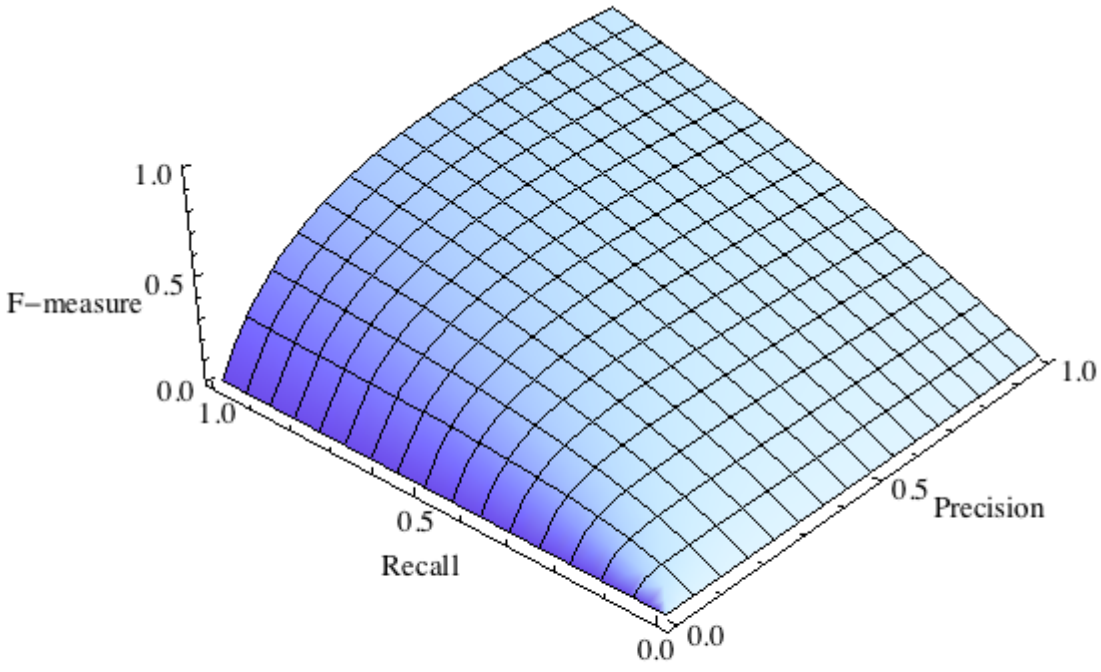
где β принимает значения в диапазоне $0 < \beta < 1$ если вы хотите отдать приоритет точности, а при $\beta > 1$ приоритет отдается полноте. При $\beta = 1$ формула сводится к предыдущей, и вы получаете сбалансированную F-меру (также ее называют F_1).



Сбалансированная F-мера



F-мера с приоритетом точности ($\beta^2 = \frac{1}{4}$)



F-мера с приоритетом полноты ($\beta^2 = 2$)

F-мера является хорошим кандидатом на формальную метрику оценки качества классификатора. Она сводит к одному числу две других основополагающих метрики: точность и полноту. Имея в своем распоряжении подобный механизм оценки вам будет гораздо проще принять решение о том являются ли изменения в алгоритме в лучшую сторону или нет.