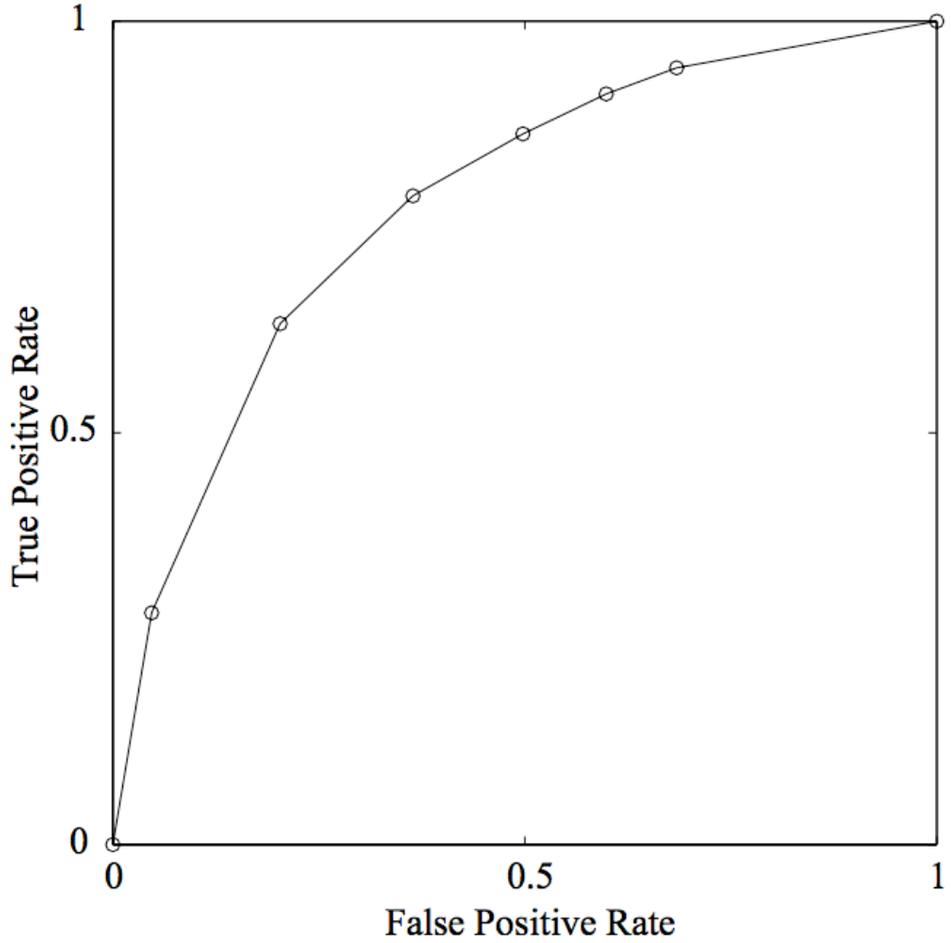
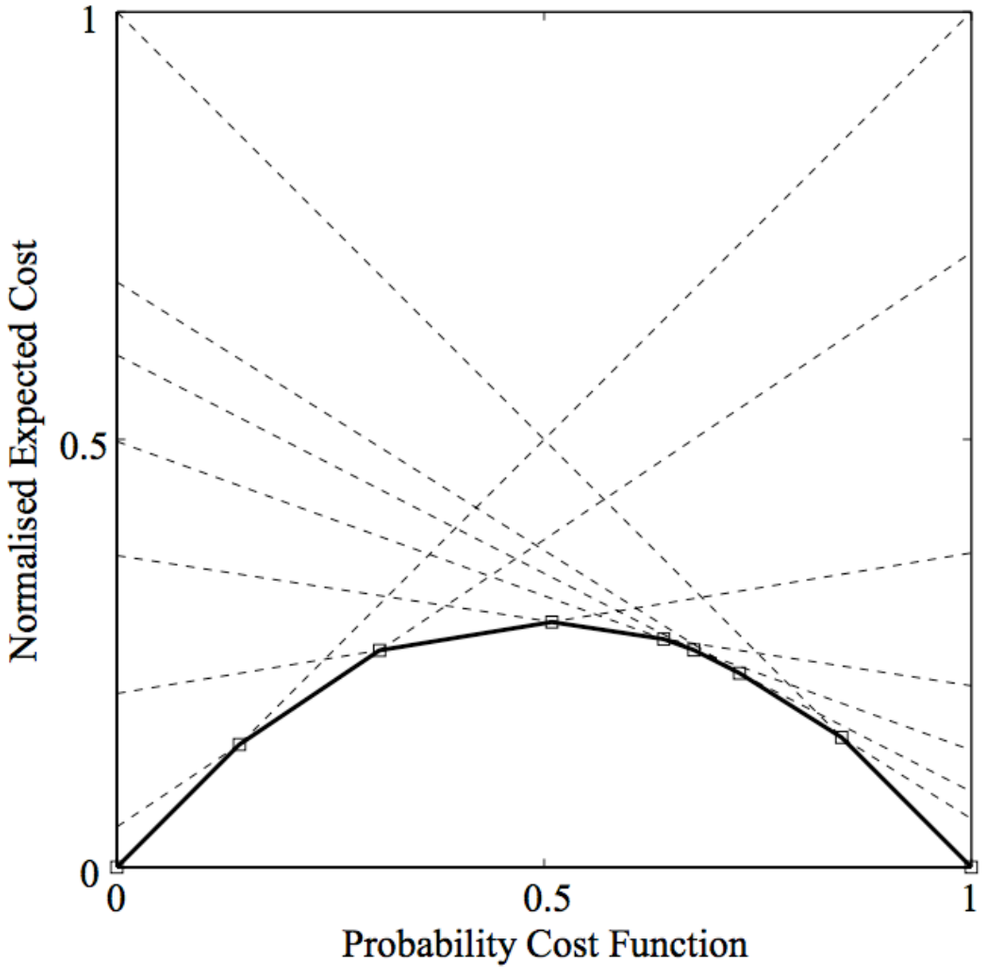
ROC, Cost, PR кривые

Гиркин Валерий
4 апреля 2017



Для оценки качества моделей бинарной классификации часто используют ROC-кривую. ROC-пространство это плоскость с осью X, отвечающей за FPR, и осью Y - за TPR. Одна матрица ошибок порождает одну точку в этом пространстве. Разные матрицы ошибок получают, например, изменением порога классификации.

ROC-кривая получается при соединении ROC-точек алгоритма с точками $(0, 0), (1, 1)$, в которых все объекты относятся к отрицательному или положительному классам соответственно.



Cost-пространством является плоскость с осью Y, отвечающей за качество (ожидаемую стоимость) модели, и с осью X, отвечающей за PCF(+):

$$PCF(+) = \frac{p(+)C(-|+)}{p(+)C(-|+) + p(-)C(+|-)}$$

При $C(-|+) = C(+|-)$ по оси x будет отложена вероятность появления положительного класса, по оси y будет отложена ошибка классификации.

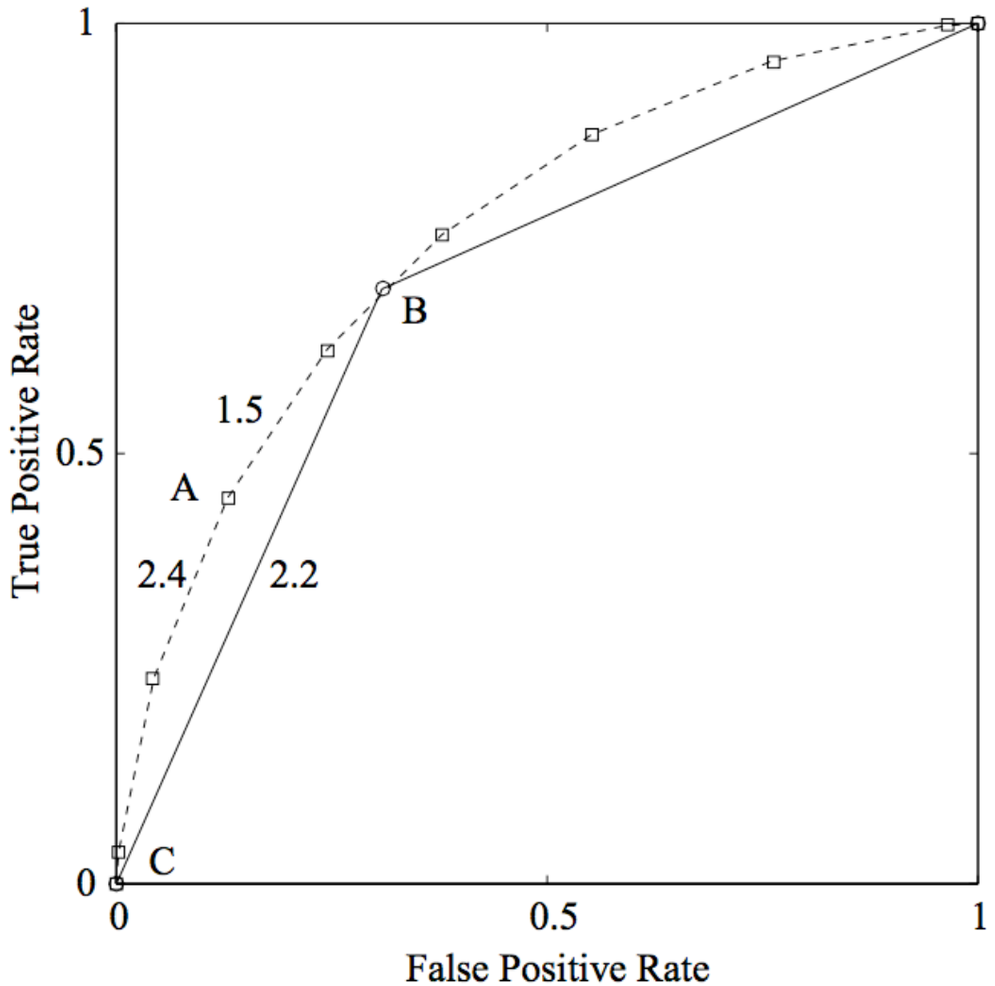
Связь ROC и Cost кривых

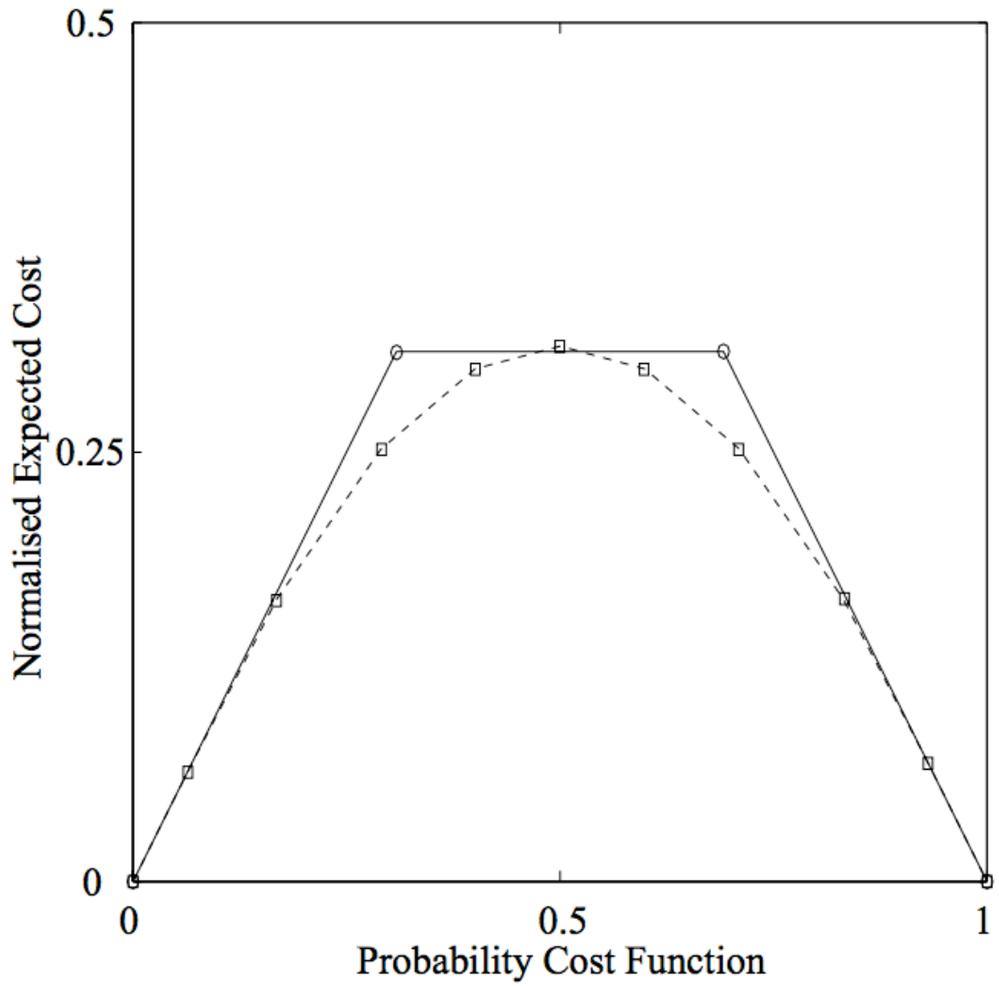
ROC и Cost пространства двойственны, т.е. точка в ROC является прямой в Cost, прямая в ROC — точка в Cost, и наоборот.
Так, точка (FP, TP) в ROC пр-ве — прямая в Cost пр-ве проходящая через точки (0, FP) и (1, 1 - TP), а её уравнение: $Y = (FN - FP) * p(+) + FP$

Прямая в ROC пр-ве с наклоном S и пересекающая ось Y в TP_0 переходит в точку в Cost пространстве по уравнениям:

$$X = p(+) = \frac{1}{1+S} \qquad Y = (1 - TP_0) * p(+)$$

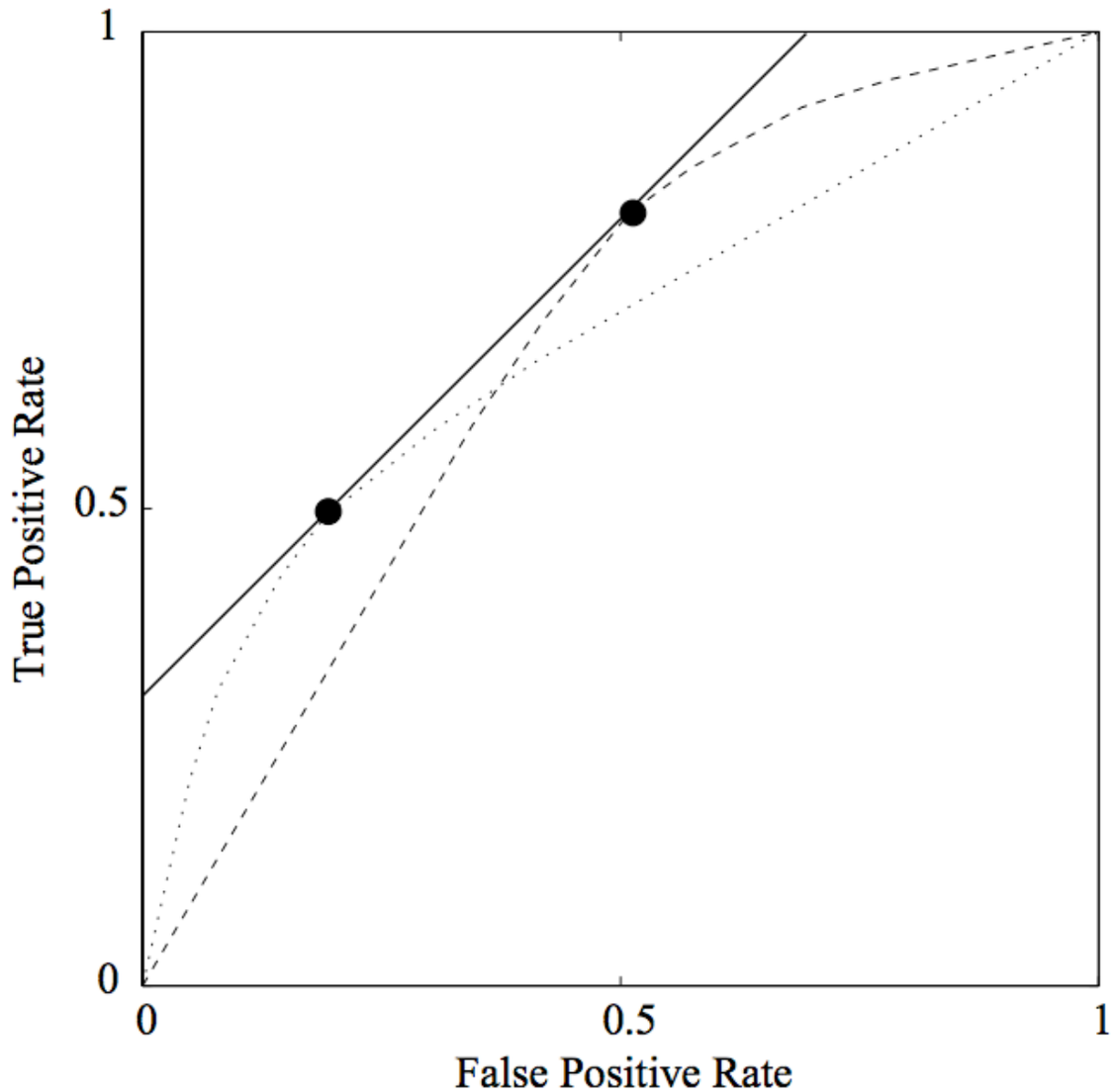
Обе эти операции обратимы. Точке (X, Y) в Cost пр-ве соответствует прямая в ROC пр-ве: $TP = (1/X - 1) * FP + (1 - Y/X)$
Прямая $aX + b$ в Cost пр-ве переходит в точку ROC пр-ва: $FP = b, TP = 1 - (b + a)$


1
0.5
0
0
0.5
1
True Positive Rate
False Positive Rate
A
B
C
1.5
2.4
2.2


0.5
0.25
0
0
0.5
1
Normalised Expected Cost
Probability Cost Function

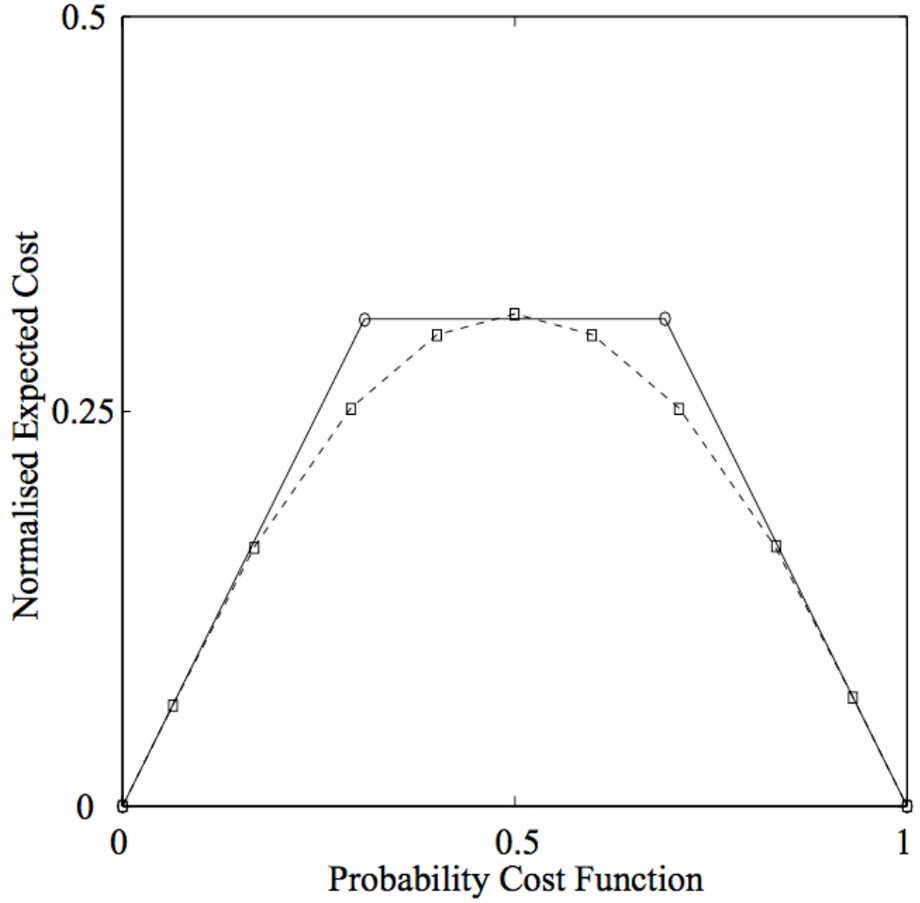
Недостатки ROC-кривых

По ROC-кривой сложно сравнить модель со случайным предсказанием, а также сложно понять, насколько одна модель лучше другой модели и понять при каких начальных условиях одна модель лучше другой.

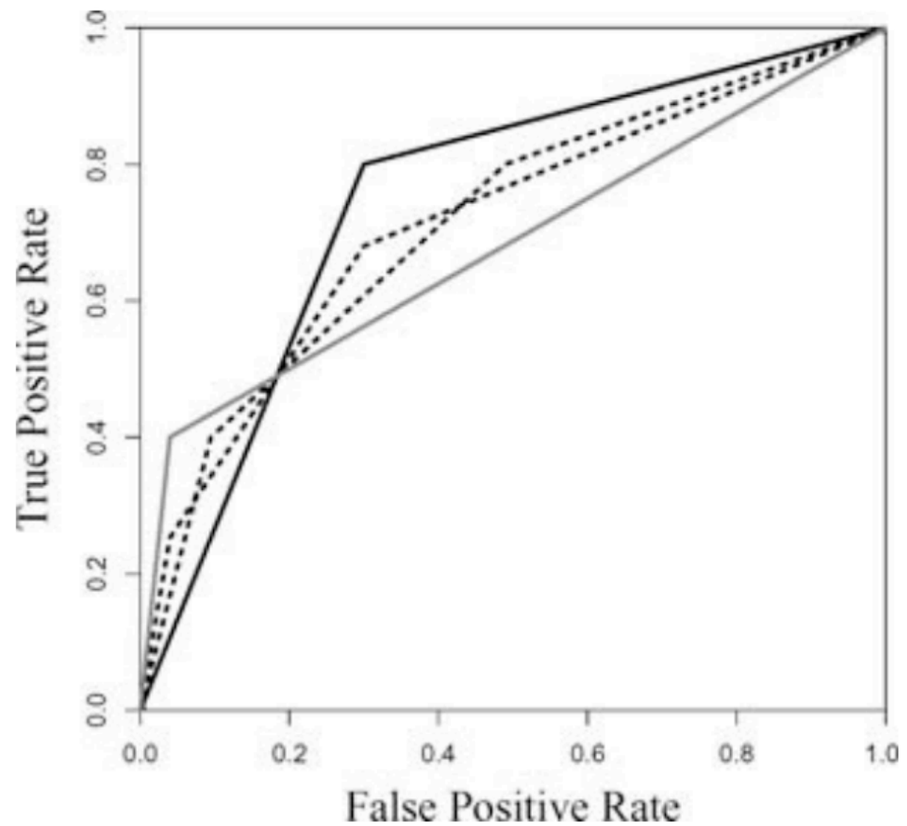


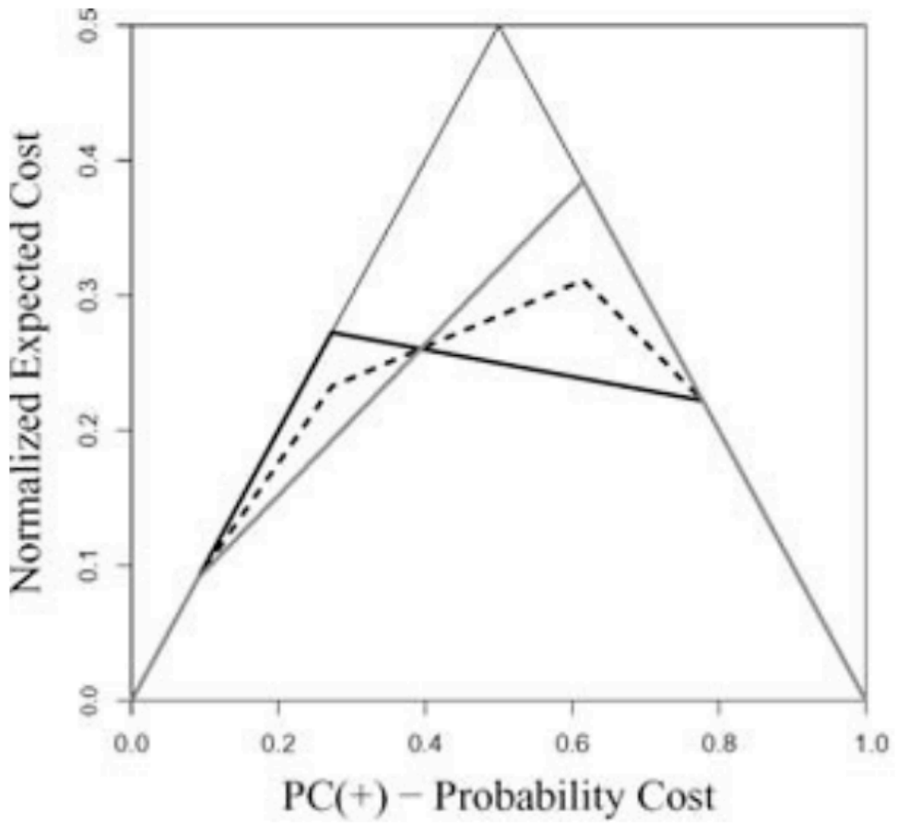
Как эти проблемы решаются в Cost пр-ве

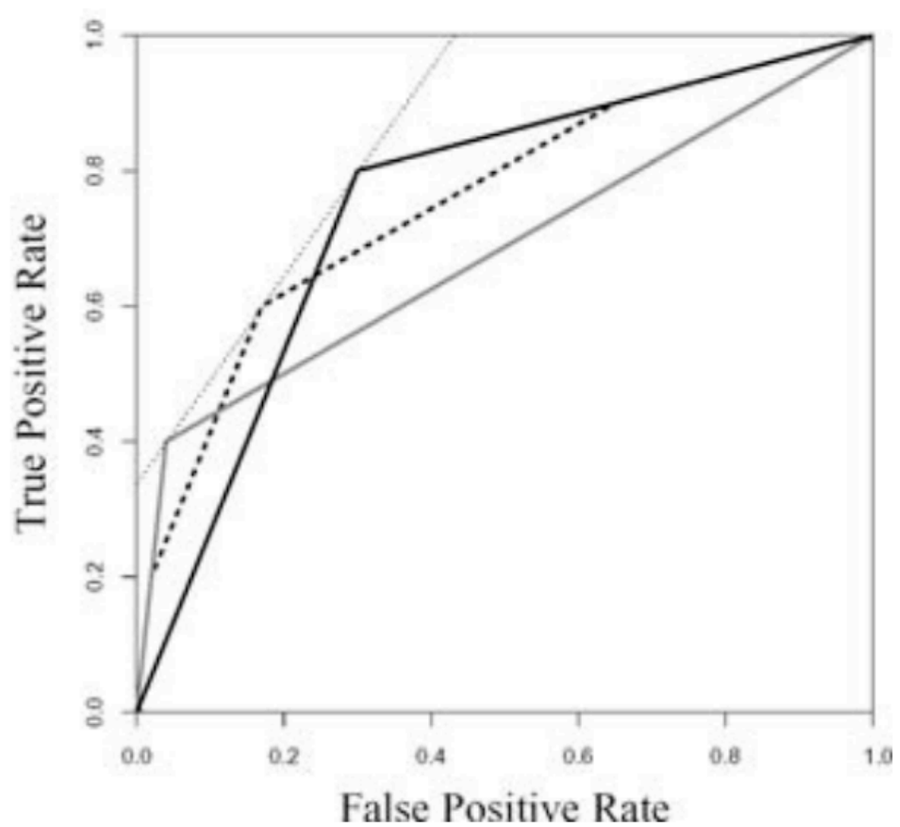
Чтобы оценить качество модели при различных начальных условиях нужно посмотреть на точку с соответствующей условию координатой X. Константные предсказания в cost-пространстве описываются диагоналями единичного квадрата. В Cost пр-ве две модели можно сравнить по разности координаты Y при одном X.



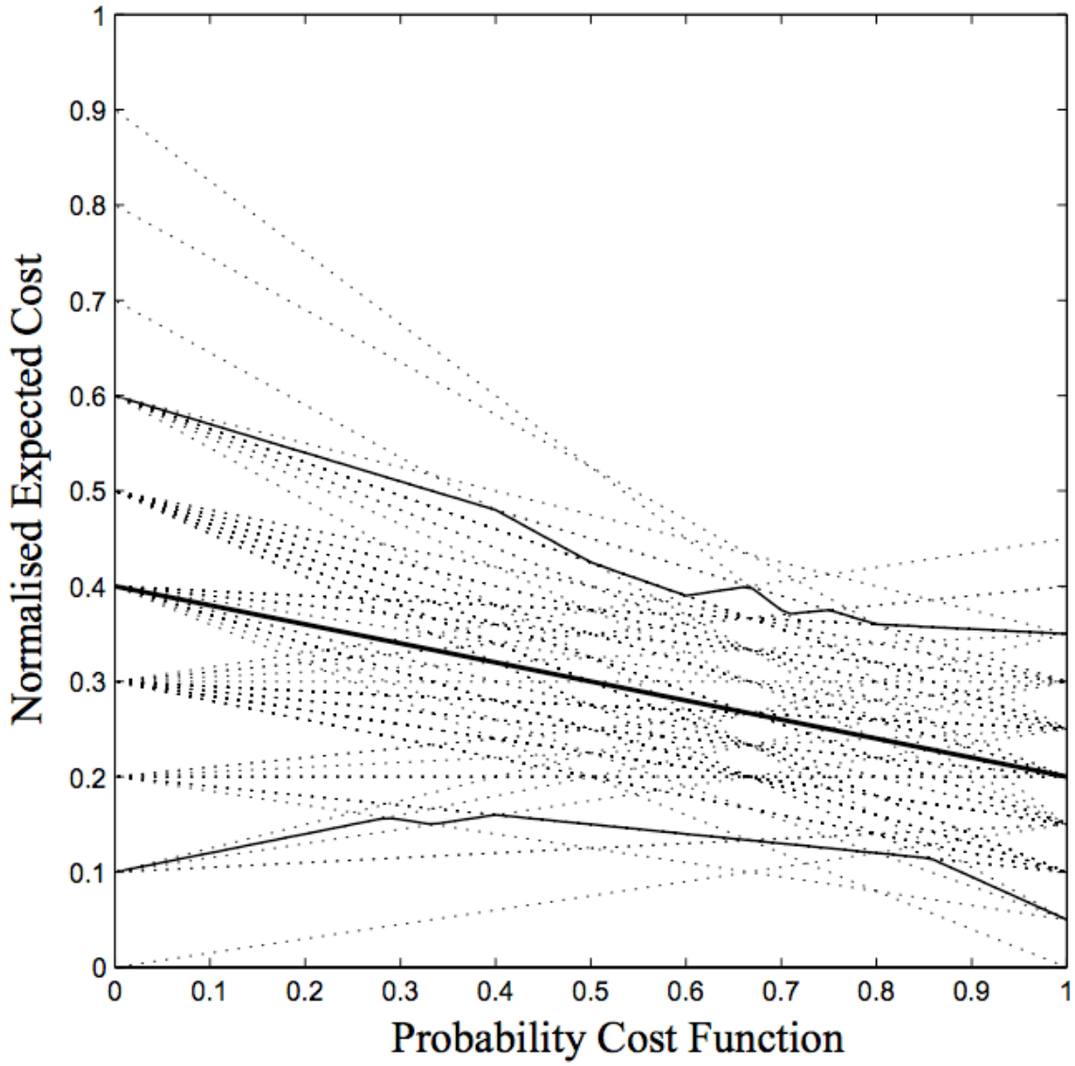
Также, в ROC пр-ве нельзя правильно усреднить несколько кривых, полученных, например, по результатам кросс-валидации, потому что усреднять нужно и FPR, и TPR. В Cost пр-ве же усреднение “по-вертикали” дает нам действительную усредненную кривую.


True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

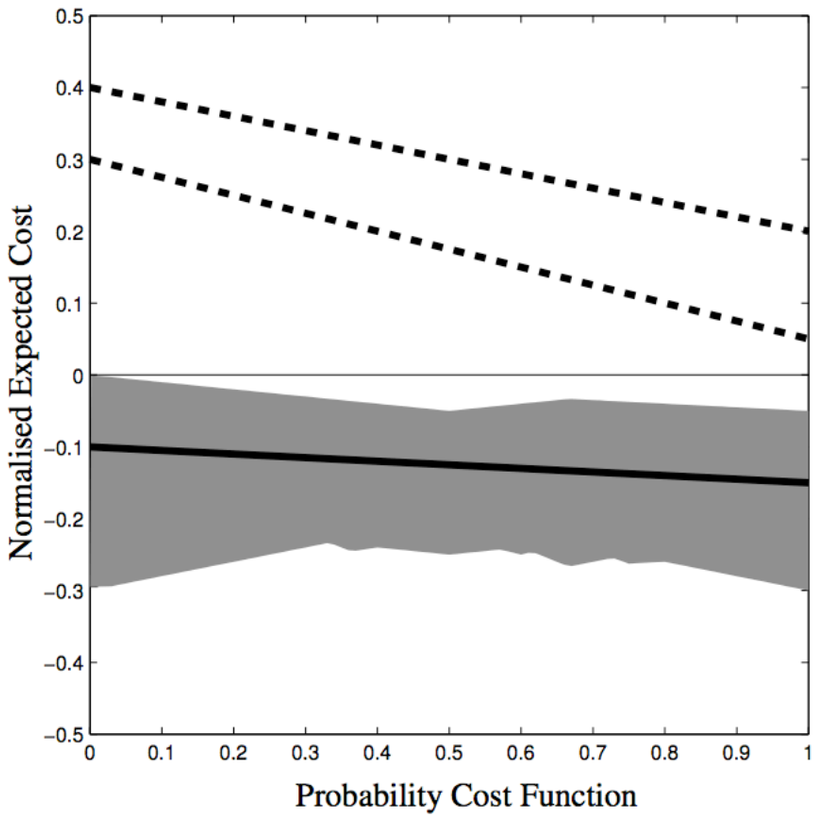

Normalized Expected Cost
PC(+) − Probability Cost
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.8
1.0

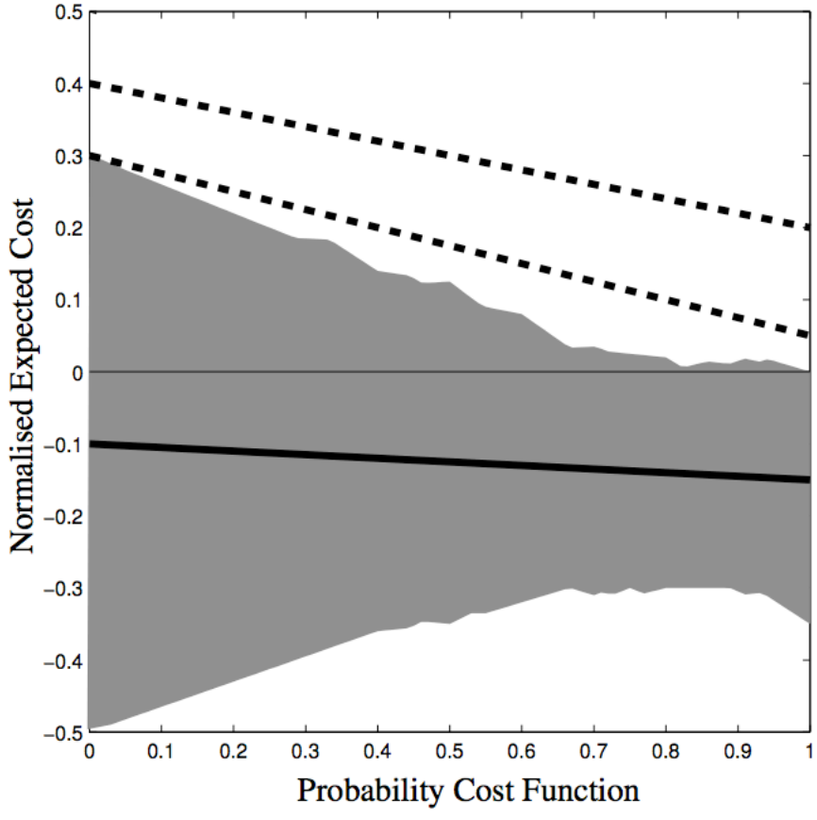

True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

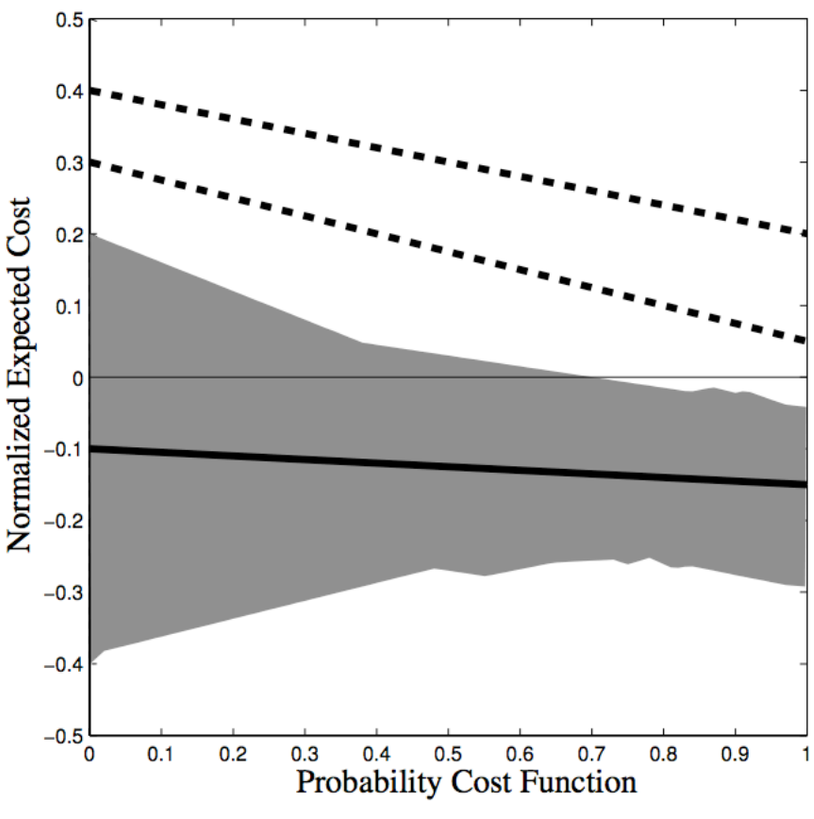
С помощью Cost-кривых и бутстрапа можно строить доверительные интервалы и проверять гипотезу о существенности различий работы двух моделей

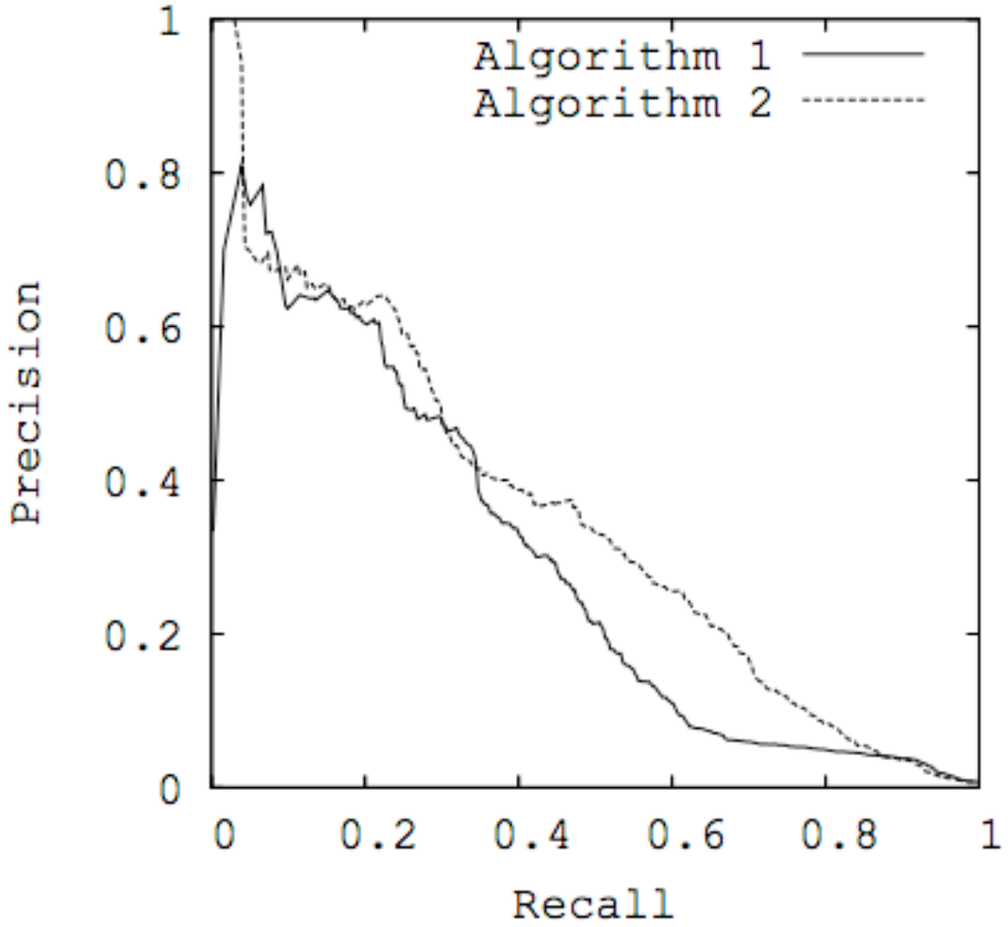


Различие в работе двух моделей является статистически значимым, если доверительный интервал разности кривых не содержит 0. Разность кривых - кривая, каждая её точка строится по разности матриц ошибок исходных кривых.

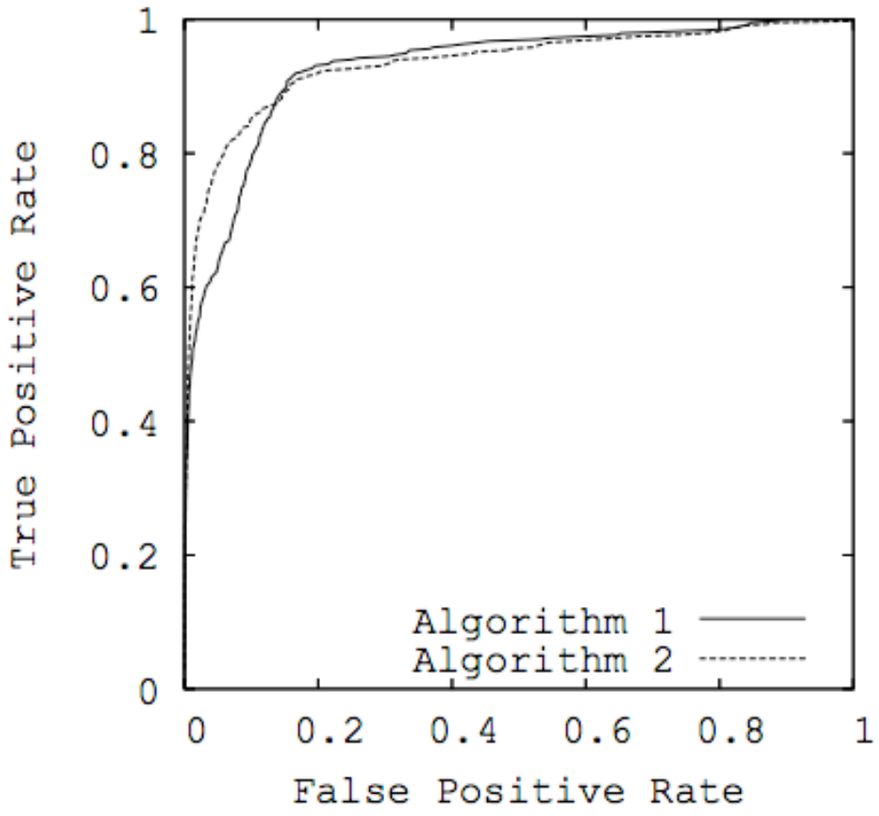




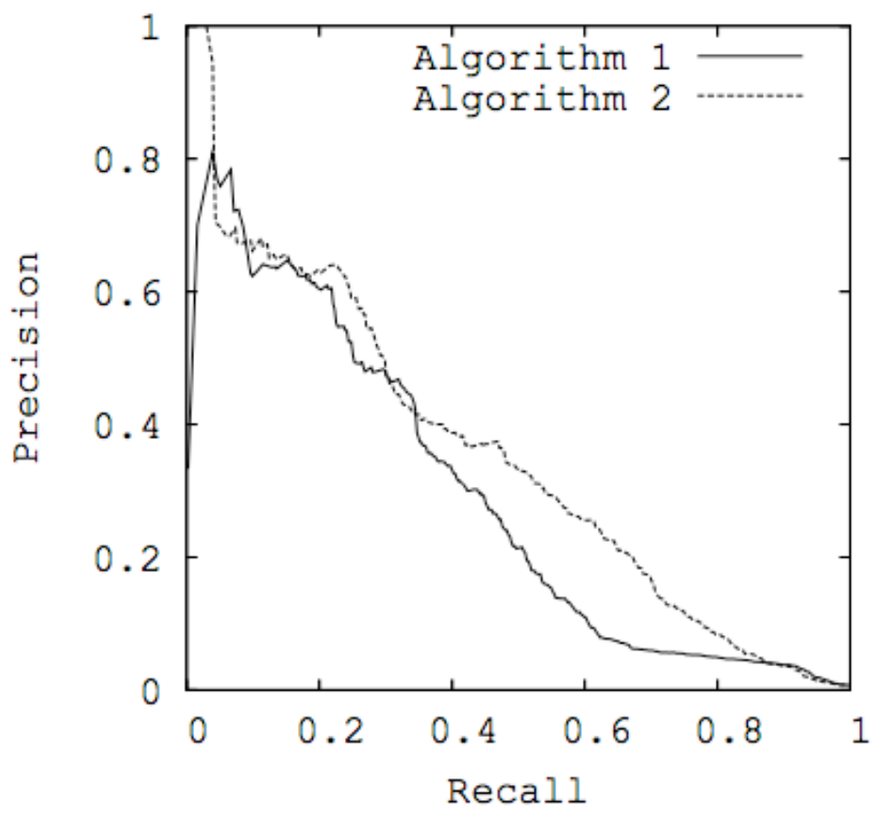




PR-пространство это плоскость с осью X, отвечающей за Recall (=TPR), и осью Y, отвечающей за Precision. PR-кривые хорошо справляются с представлением моделей, обученных на сильно смещенных данных.



(a) Comparison in ROC space



(b) Comparison in PR space

Связь ROC и PR кривых

Если данные фиксированны, то по точке в ROC пр-ве можно однозначно восстановить матрицу ошибок, а по ней можно найти соответствующую точку в PR пр-ве. Если Recall не равен 0, то то же самое можно сделать и в обратную сторону. Поэтому между этими кривыми есть взаимно-однозначное соответствие.

На фиксированном наборе данных одна кривая в ROC пр-ве доминирует над второй тогда и только тогда, когда первая доминирует над второй в PR пр-ве.

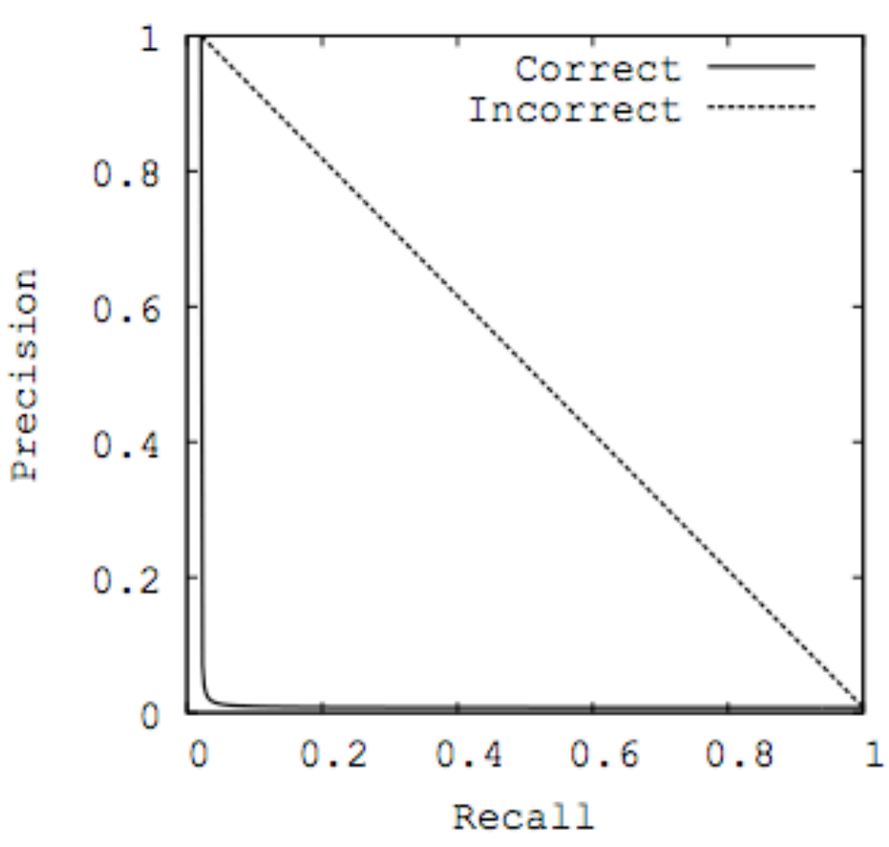
Выпуклая оболочка и AUC

Выпуклая оболочка ROC-кривой может быть построена линейной интерполяцией точек. В PR пр-ве это не работает, потому что таким образом мы будем увеличивать AUC-PR. Достижимая PR-кривая доминирует все возможные PR-кривые, построенные на тех же точках. Она строится по точкам выпуклой оболочки соответствующей ROC-кривой.

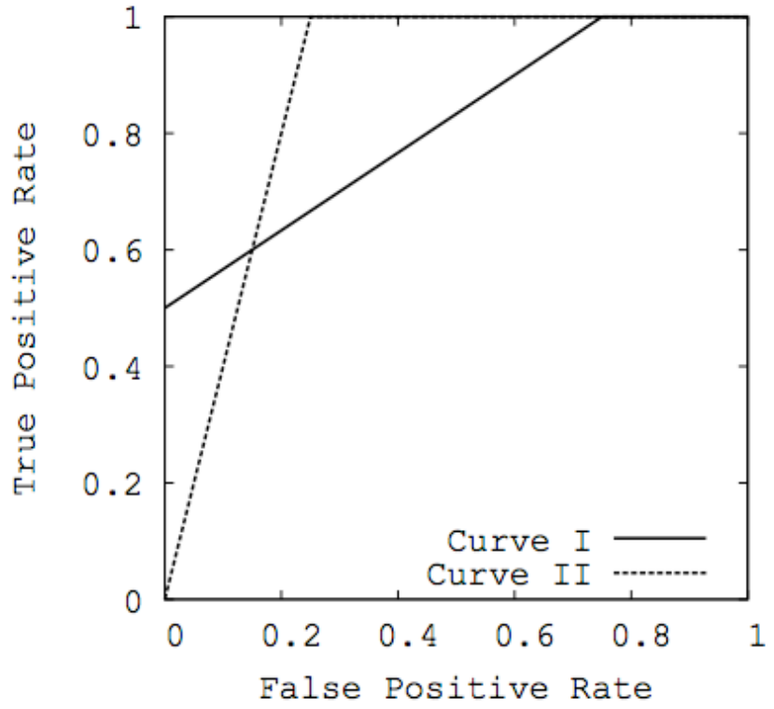
Пусть точки A и B соседние на ROC-кривой, тогда соответсвующий участок AB PR-кривой строится по следующему правилу:

$$\left(\frac{TP_A + x}{TotalPos}, \frac{TP_A + x}{TP_A + x + FP_A + \frac{FP_B - FP_A}{TP_B - TP_A} x} \right), x \in \{1, \ldots, TP_B - TP_A\}$$

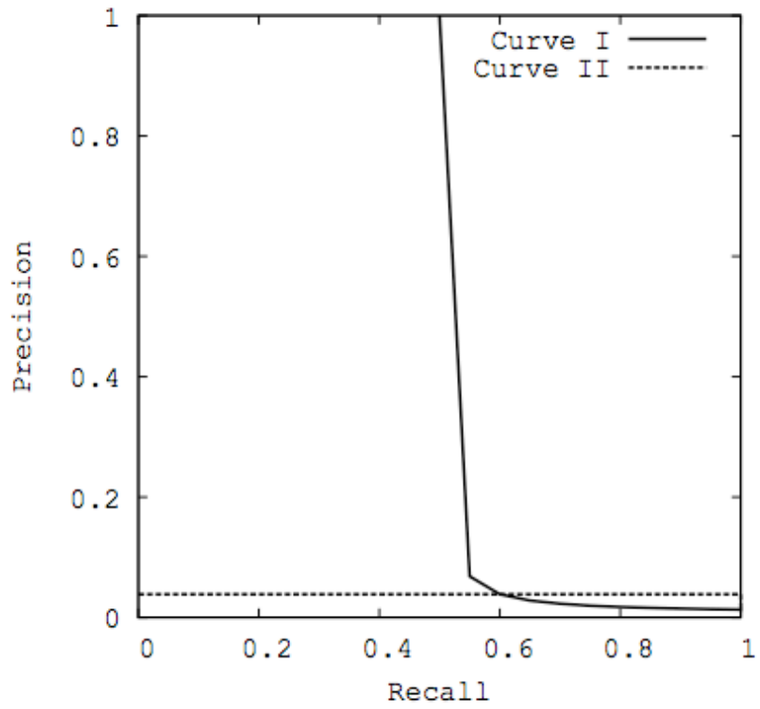
	TP	FP	REC	PREC
A	5	5	0.25	0.500
.	6	10	0.30	0.375
.	7	15	0.35	0.318
.	8	20	0.40	0.286
.	9	25	0.45	0.265
B	10	30	0.50	0.250



Оптимизация AUC



(a) Comparing AUC-ROC for two algorithms



(b) Comparing AUC-PR for two algorithms

Оптимизация AUC-ROC не всегда оптимизирует AUC-PR, что особенно видно на смещенных данных.
20 +, 2000 -: AUC-ROC I = 0.813, AUC-ROC II = 0.875. Но AUC-PR I = 0.514, а AUC-PR II = 0.038