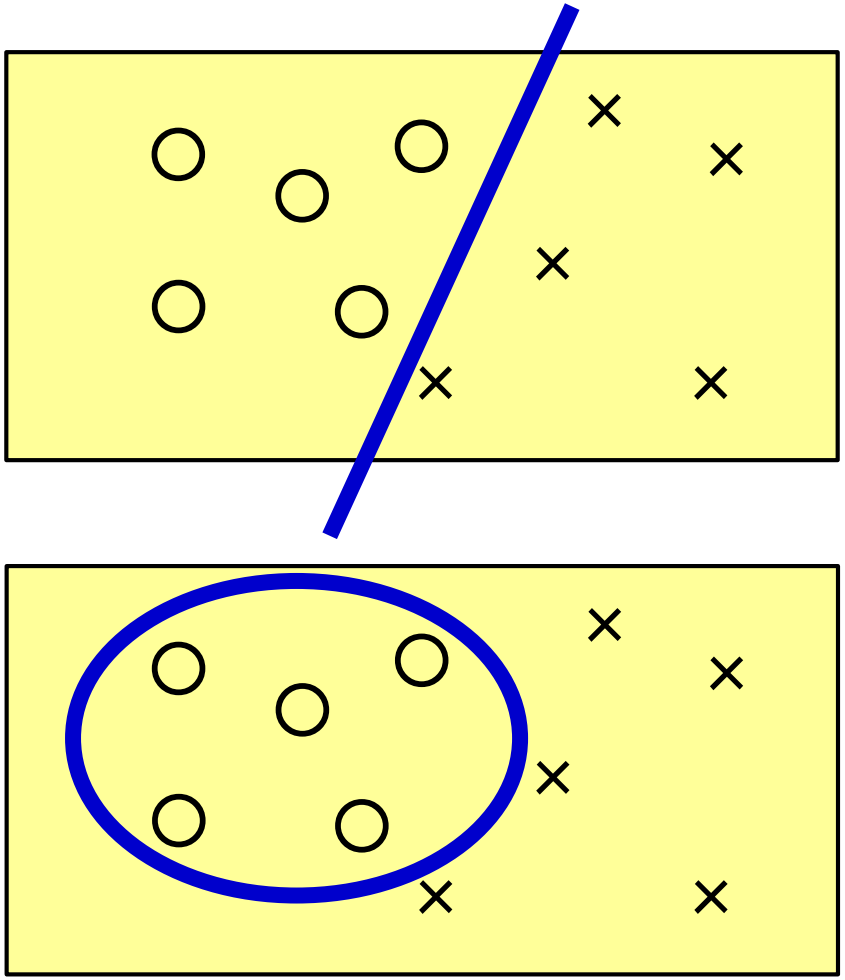
# 共生社会特論

## 第5回　評価手法

2016年1月17日

# 「分かる」←「分ける」

- 「分かる」とは区別できること

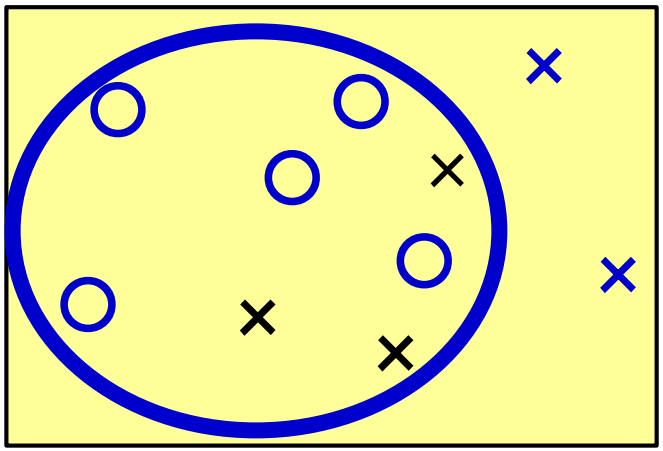


識別器
(discrminator)
二項分類器
(binary classifier)

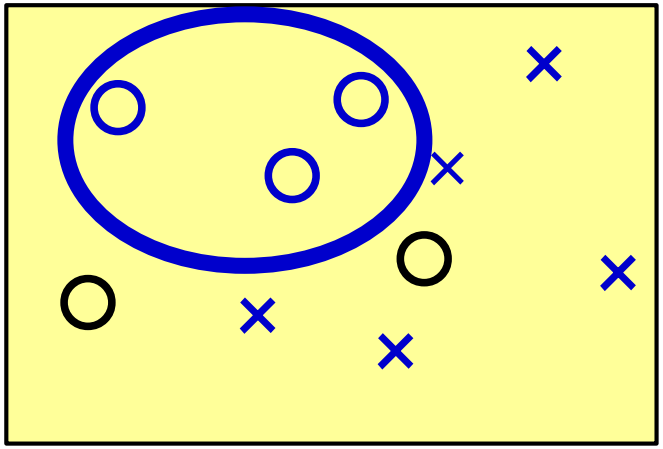
# 正解率 (Accuracy)

- 二項分類器の性能を評価

$$\text{Accuracy} = \frac{\text{正しく識別できた要素数}}{\text{全要素数}}$$



$$A = \frac{7}{10}$$



$$A = \frac{8}{10}$$

# 二項分類器の性能評価

| 真の結果 / 出力結果 | Positive | Negative |
|---|---|---|
| Positive | True Positive | False Positive (Type I error) |
| Negative | False Negative (Type II error) | True Negative |

$$\text{Accucary} = \frac{tp+tn}{tp+tn+fp+fn}$$

註：真の結果の代わりに Gold Standard のことも

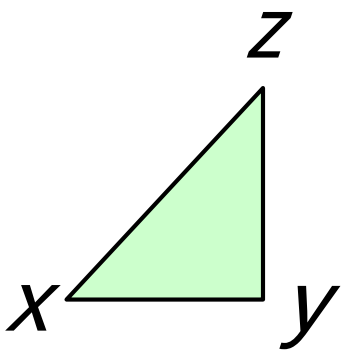
# 情報検索 (Information Retrieval)

- 二項分類の一種
- 大量の文書群中から求めるものを抽出
  - 一般に、入力キーワードと類似する文書を抽出
    - 類似度が閾値 $\theta$ 以上
- 論文検索、Google 検索など

- 正解率が性能評価に適さない

# 類似度と距離

- 類似度 (similarity)
  - 値が大きいほど似ている
  - 負の値をとることもある（例：コサイン類似度）
- 距離 (distance)
  - 値が小さいほど似ている
  - 距離の公理を満たす必要がある
    - $d(x,y) \geqq 0$
    - $x = y \Leftrightarrow d(x,y) = 0$
    - $d(x,y) = d(y,x)$
    - $d(x,y) + d(y,z) \geqq d(x,z)$



# 情報検索と正解率

| 真の結果 / 出力結果 | Positive | Negative |
|---|---|---|
| Positive | 100 | 20 |
| Negative | 30 | 100,000,000 |

$$\text{Accucary} = \frac{tp+tn}{tp+tn+fp+fn} = \frac{100{,}000{,}100}{100{,}000{,}150}$$

$fn$ が大きすぎて評価できない

# 情報検索に適した評価指標

| 真の結果 / 出力結果 | Positive | Negative |
|---|---|---|
| Positive | 100 | 20 |
| Negative | 30 | 100,000,000 |

精度

$$\text{Precision} = \frac{tp}{tp+fp}$$

$$\frac{100}{100+20}$$

再現率

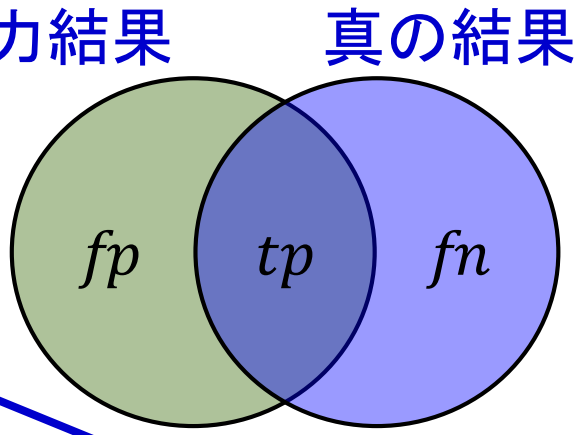
$$\text{Recall} = \frac{tp}{tp+fn}$$

$$\frac{100}{100+30}$$

# 精度と再現率

- 精度（適合率）
  - 出力結果の内、正しい割合

$$\text{Precision} = \frac{tp}{tp+fp}$$




分子は同じ

- 再現率
  - 求めるモノの内、抽出できた割合
  - 計算が困難な場合も

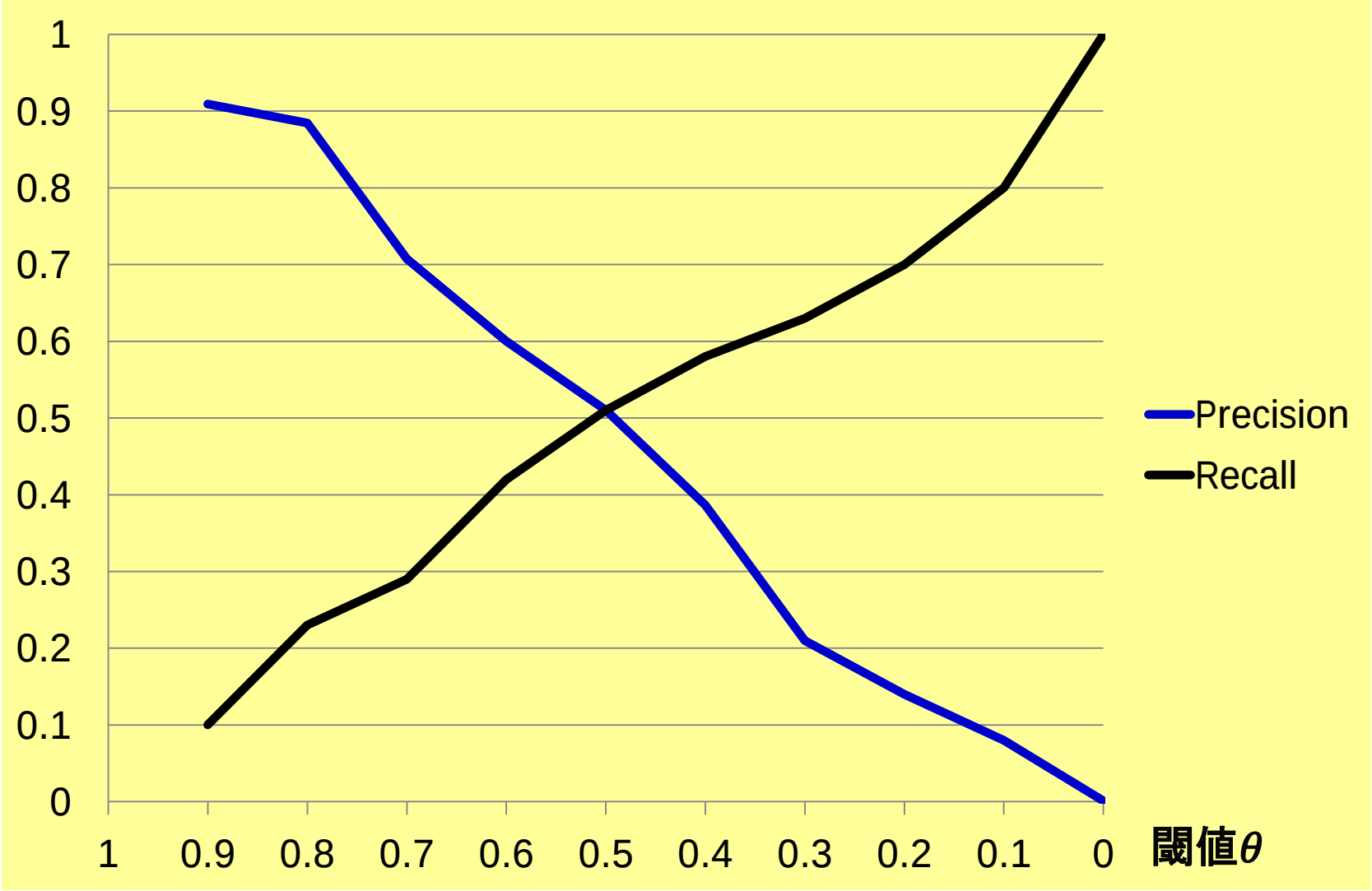
$$\text{Recall} = \frac{tp}{tp+fn}$$

# トレードオフ (Trade-off)

- 再現率100%は簡単
  - 全文書を出力する（$\theta = 0$）
  - 実用上は役に立たない
- 精度100%
  - 類似度が最大となる1個だけ出力
  - 漏れが多くなる

- いずれを重視するかは目的次第
  - 再現率重視: 特許検索、論文のサーベイ
  - 精度重視: 一般の検索エンジン

# 精度と再現率の変化


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1 0.9 0.8 0.7 0.6 0.5 0.4 0.3 0.2 0.1 0
閾値θ
Precision
Recall

# 精度と再現率の統合

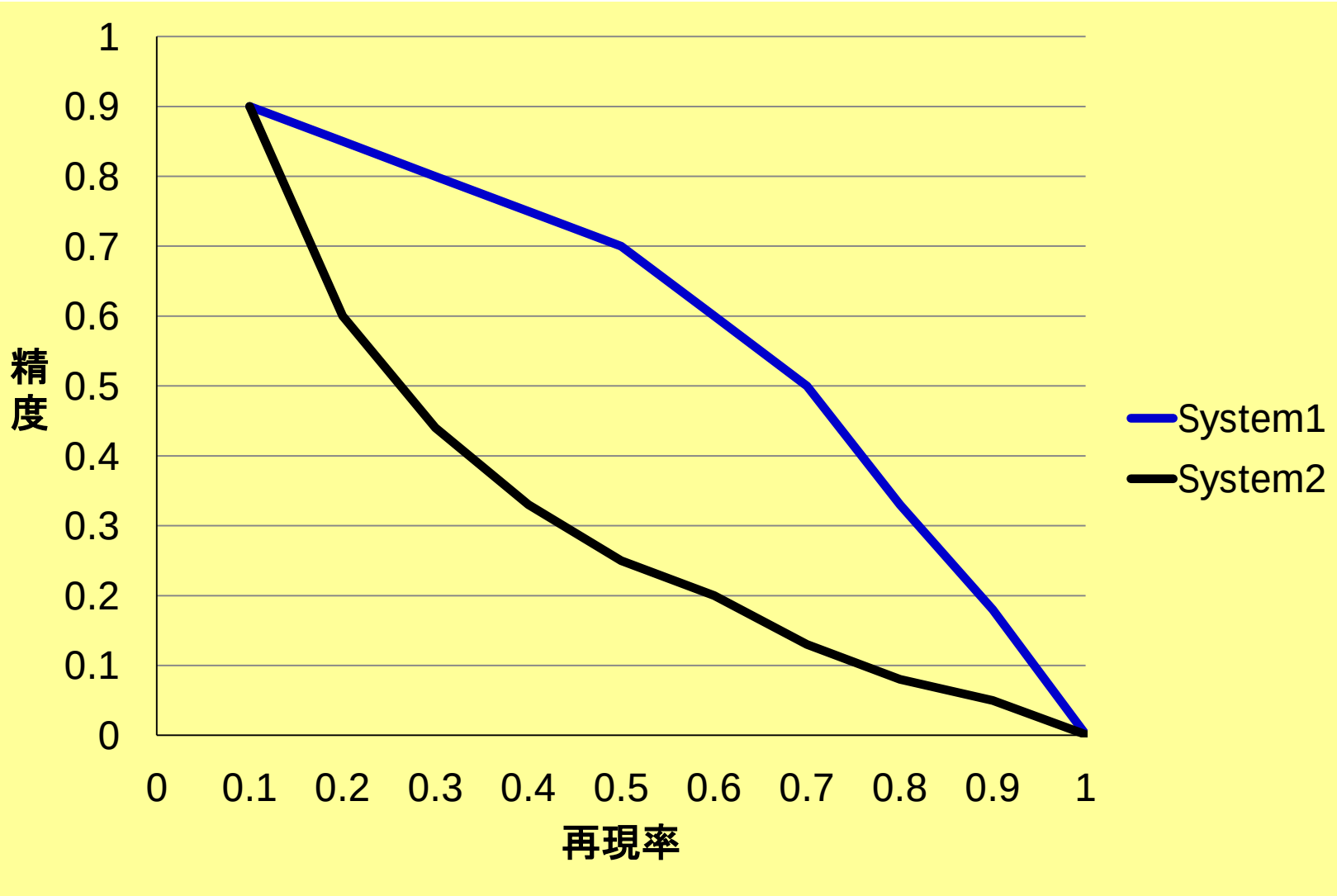
- F値 (F-measure)
  - 精度と再現率の調和平均

$$F = 2\frac{\text{Precison} \times \text{Recall}}{\text{Precison} + \text{Recall}}$$

- Precision-recall breakeven point
  - 精度と再現率が等しくなったときの値

- 11点平均精度 (11-pt average precision)
  - 再現率がそれぞれ 0.0, 0.1, …, 1.0 となる11点における精度の平均（通常は補完する）

# 精度-再現率曲線


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
精度
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
再現率
System1
System2

# オマケ：医療用語

| 真の状態 / 検索結果 | Positive | Negative |
|---|---|---|
| Positive | True Positive | 偽陽性 |
| Negative | 偽陰性 | True Negative |

再現率＝感度 $\text{Sensitivity} = \frac{tp}{tp+fn}$

特異度 $\text{Specificity} = \frac{tn}{tn+fp}$

ROC曲線

註：偽陽性≠擬陽性

# 問題

- 癌検診の結果が「要再検査」＝positive
- 癌の人が「要再検査」になる確率
  ＝感度（再現率） 90%
- 癌でない人が「要再検査」になる確率
  ＝特異度 10%
- 検診を受ける人の内、癌の人の割合 0.1%
- 「要再検査」の人が本当に癌の確率
  ＝精度は？

# 答

1万人が受診と仮定

| 検索結果＼真の状態 | Positive | Negative |
|---|---|---|
| Positive | 9 | 999 |
| Negative | 1 | 8991 |

$$\text{再現率＝感度 Sensitivity} = \frac{tp}{tp+fn} = \frac{9}{9+1} = 0.9$$

$$\text{特異度 Specificity} = \frac{tn}{tn+fp} = \frac{999}{8991+999} = 0.1$$

$$\text{精度 Precision} = \frac{tp}{tp+fp} = \frac{9}{9+999} = 0.0089$$

# 解説

- 「検診を受ける人の内、癌の人の割合」が重要なパラメータ
  - 0.1%→10%の場合

| 検索結果＼真の状態 | Positive | Negative |
|---|---|---|
| Positive | 900 | 900 |
| Negative | 100 | 8100 |

$$\text{精度 Precision} = \frac{tp}{tp+fp} = \frac{900}{900+900} = 0.5$$

# 現実問題への適用

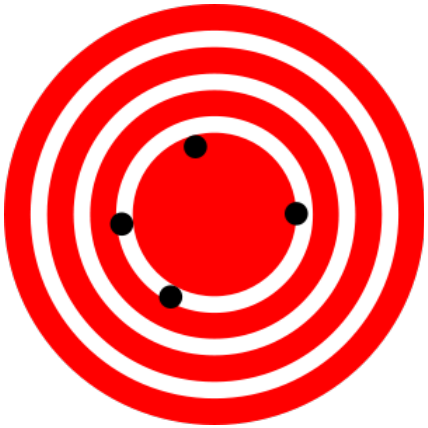
| 真の結果 / 出力結果 | Positive | Negative |
|---|---|---|
| Positive | True Positive | False Positive |
| Negative | False Negative | True Negative |

- 完璧な検査（精度と再現率が100%）はない
- 目的により重視するものが変わる
  - 癌検査、冤罪、不正受給

# 正確度と精度の異なる用法

測定などでは意味が異なる

- 正確度 (accuracy)
  - 真値との近さを示す尺度
- 精度 (precision)
  - 複数回の値の間でのばらつきの尺度



高正確度だが低精度

高精度だが、低正確度

# 実験と評価

# 分類器の構築と評価

訓練データ (training set) で学習後、
テストデータ (test set) で評価

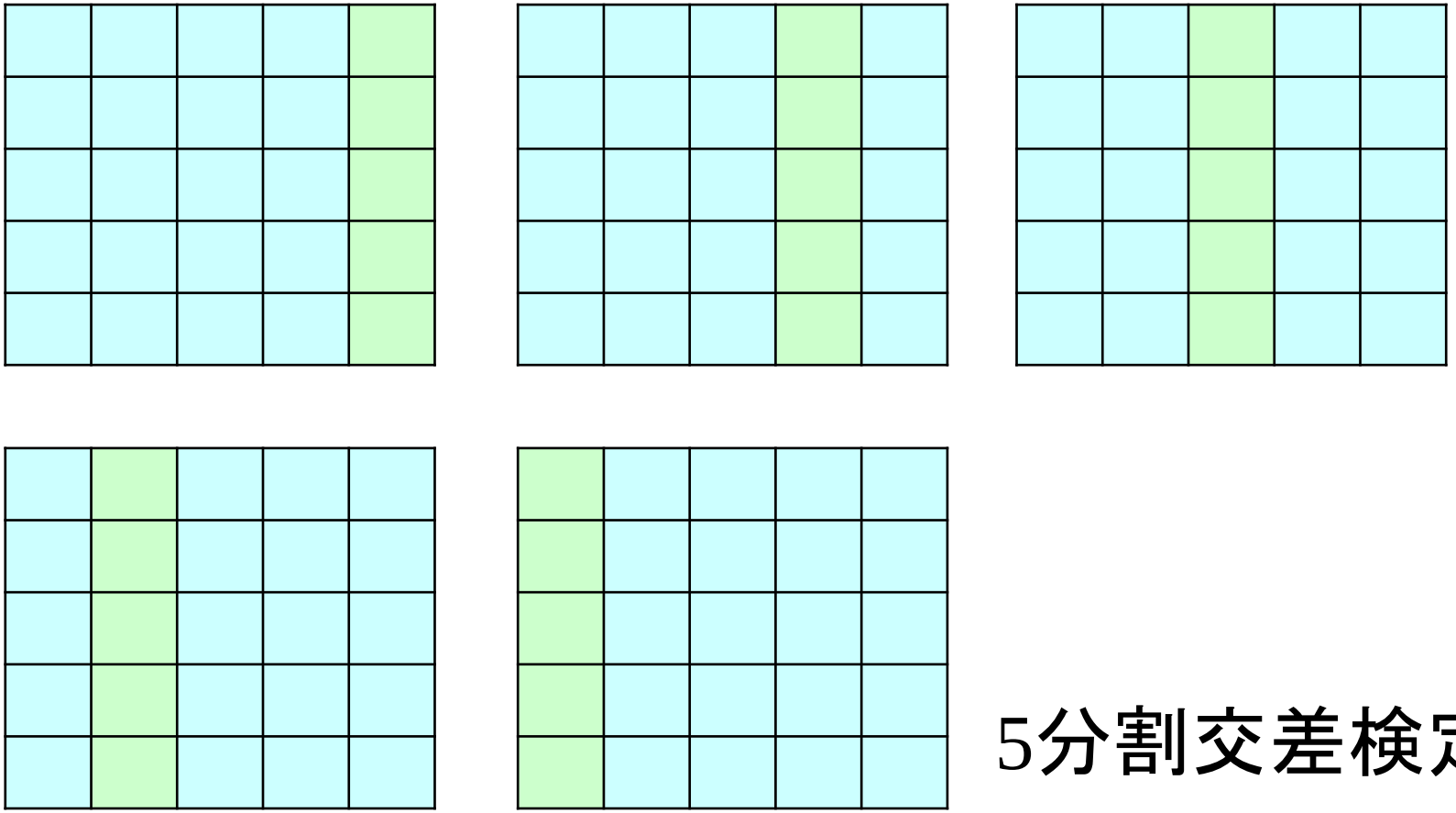
- Closed test
  - 訓練データとテストデータが同じ
- Open test
  - 訓練データとテストデータが異なる

どちらのデータも正解が必要

大量に用意できない

# k-分割交差検定 (k-fold cross-validation)

- 訓練データ と テストデータ を交替して実験



5分割交差検定

# 検定

# 母集団と標本

- 母集団 (population)
  - 調査したい対象全体の集合
  - 母平均 $\mu$
  - 母分散 $\sigma^2$
- 標本 (sample)
  - 母集団から無作為抽出した実際の調査対象
  - 標本調査を複数回することもある
  - 標本サイズ $n$
  - 標本平均 $\bar{x}$
  - 標本分散 $s^2$

# 誤用

- 「母数」(parameter)
  - 母集団の特徴を表す特性値
  - 「母平均」「母分散」など
  - 「分母」や「全数」のことではない

- 「標本数」(sample size)
  - 意味が曖昧なのでこの訳は避けるべき
  - 「標本の大きさ」か「標本の個数」か不分明
  - 「標本サイズ」「標本の大きさ」が望ましい

# 平均

- 相加平均(算術平均)

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- 相乗平均(幾何平均)

$$\sqrt[n]{\prod_{i=1}^{n} x_i} = \sqrt[n]{x_1 x_2 \cdots x_n}$$

- 調和平均

$$\frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

# マイクロ平均とマクロ平均

| キーワード | apple | banana | cherry | durian | eggfruit |
|---|---|---|---|---|---|
| 出力数 | 700 | 400 | 500 | 40 | 3 |
| 正解数 | 350 | 340 | 400 | 38 | 3 |
| 精度 | 50.0% | 85.0% | 80.0% | 95.0% | 100% |

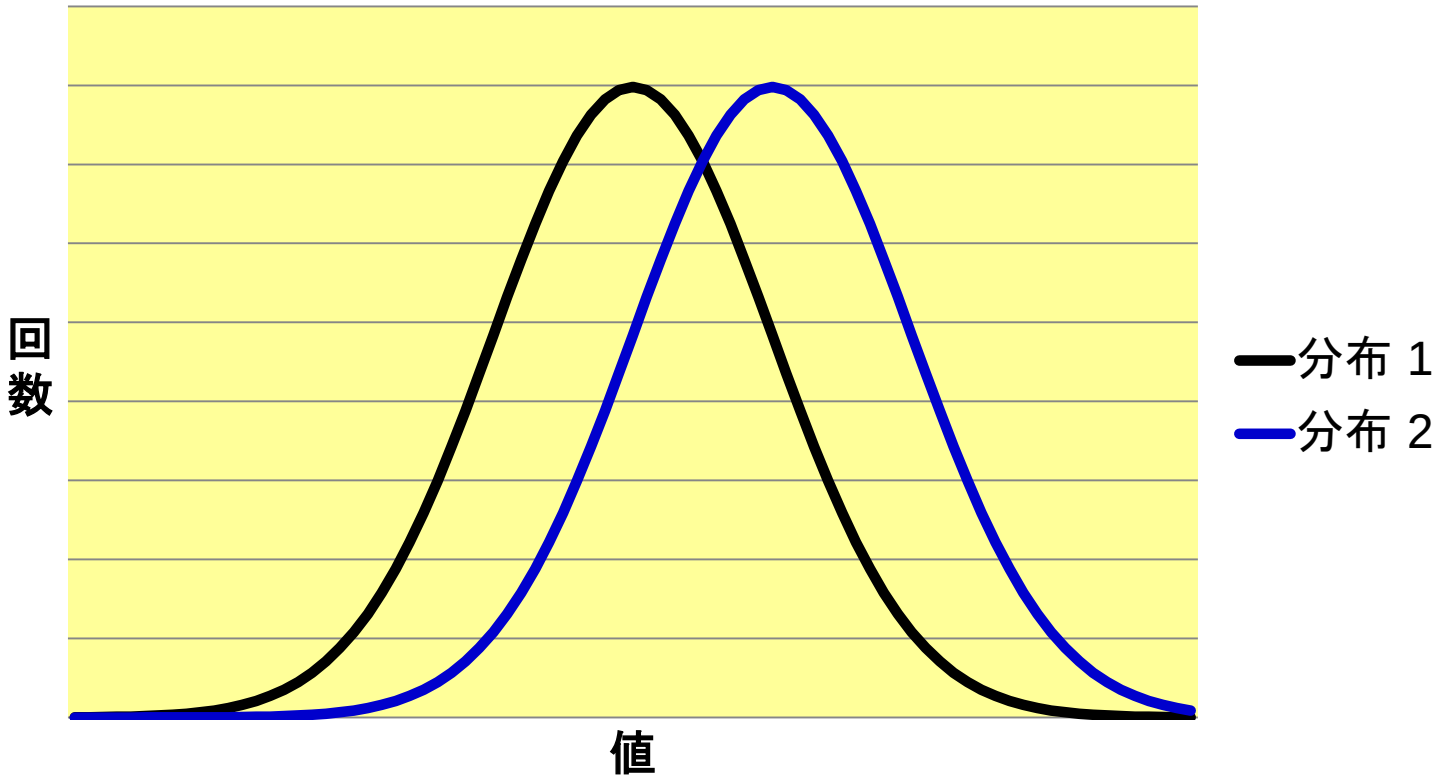
- マクロ平均

$$\frac{1}{5}\left(\frac{350}{700}+\frac{340}{400}+\frac{400}{500}+\frac{38}{40}+\frac{3}{3}\right)=0.82$$

- マイクロ平均

$$\frac{350+340+400+38+3}{700+400+500+40+3}=0.69$$

# 度数分布




この二つの分布に差はあるか？

# 有意差検定 (Significance Test)

- 差が偶然でないことの検証
- 有意水準$\alpha$の仮説検定
  - $\alpha$の値は 0.05や0.01
- これをしない限り、「有意な (significantly)」という用語を使用してはならない

# 検定の種類

- パラメトリックな検定
  - t 検定 （正規分布・同じ分散を仮定）
  - 対応のある t 検定
  - F 検定
- ノンパラメトリックな検定
  - 符号検定
  - ウィルコクソンの符号順位検定
  - マン・ホイットニーのU検定

# 仮説検定 (Statistical hypothesis test)

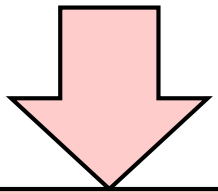
- 帰無仮説を棄却する形で確率的に判断
  - ➢ 有意差があることを示したい場合
  1. 「差がない」という帰無仮説を設定
  2. 統計量を算出
  3. 求められた統計量が起こる確率を導出
  4. 確率が$\alpha$未満なら帰無仮説を誤りとして棄却
     ⇒「差がない」が棄却されたので「差がある」

# 誤りの種類

- 第1種過誤 (Type I error)
  - 正しい帰無仮説を棄却してしまう
    - 本当は「差がない」のに「差がある」と判断
  - 起きる確率を危険率と呼ぶ
    - 有意水準$\alpha$と等しい
- 第2種過誤 (Type II error)
  - 誤った帰無仮説を棄却しない
    - 本当は「差がある」のに「差がない」と判断
  - 起きる確率を$\beta$で表現する
    - $1-\beta$を検出力(power)と呼ぶ

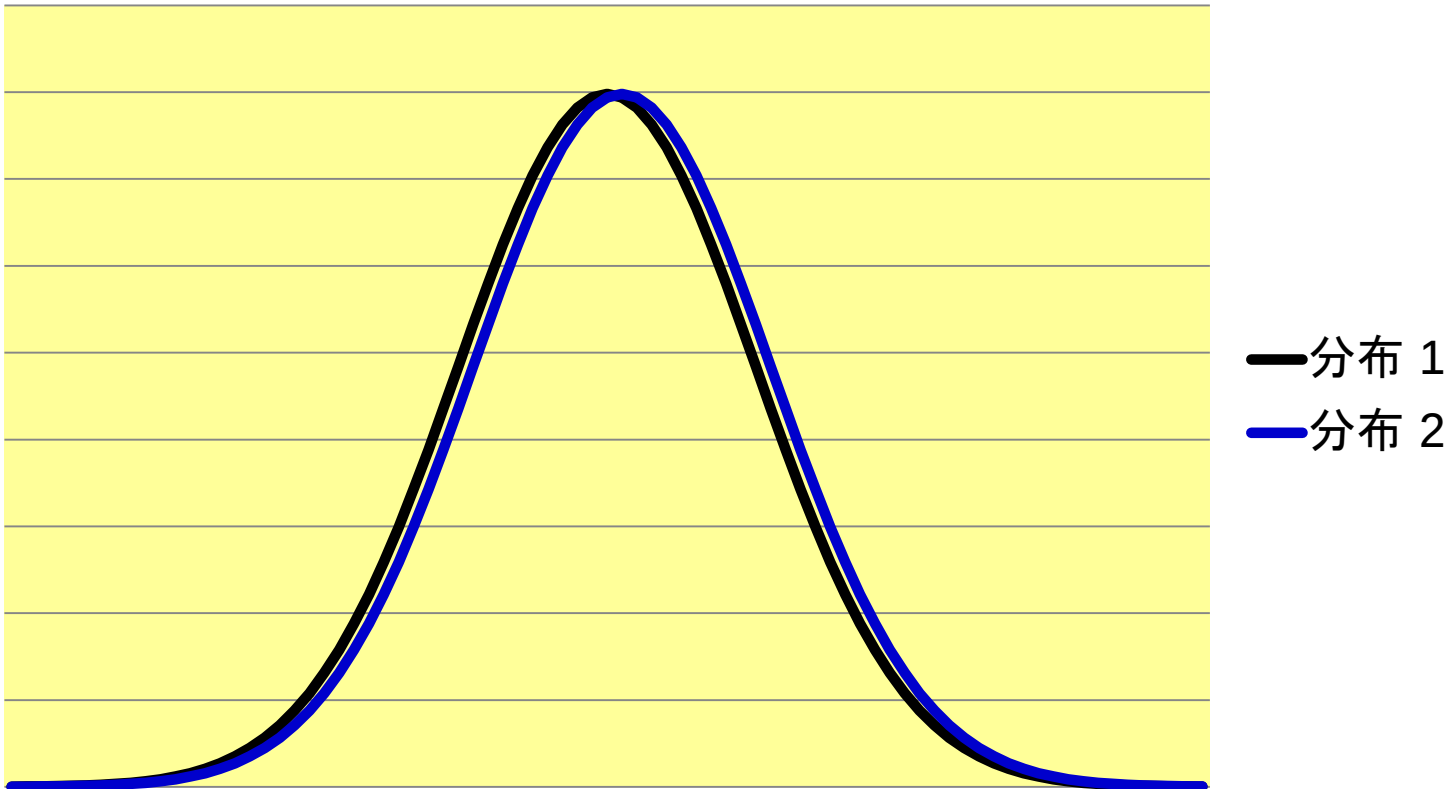
# 危険率と検出力

- 第1種過誤と第2種過誤の間はトレードオフ
  - 危険率を決め、その中の検出力最大を選択
- 多重比較の場合は$\alpha$の補正が必要
  - 例：血液型性格診断
- 標本サイズが大きくなれば検出力が上がる



標本サイズを増やせば、
どんな有意水準でもクリア可

# 役に立たない「有意差」




標本サイズを増やせば、上記でも
「有意差あり」と判断される

# 効果量 (effect size)

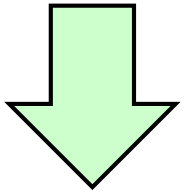
- 有意差検定は「差が偶然ではない」を判定
- 効果量は「差がどのくらいか」を判定
- t検定の場合

$$d = \frac{\bar{x} - \bar{y}}{s}$$

  - 0.20 で効果量小
  - 0.50 で効果量中
  - 0.80 で効果量大

# 有意差と効果量

標本サイズが小さいと有意差なしだが、
標本サイズを増やすと有意差あり



多くの場合効果量小

このような場合、高確率で効果が小さい

# 翻訳の自動評価

# 翻訳の評価

- 人手による評価
  - 高コスト
    - 両言語の分かる専門家
  - 基準が一定でない
  - 量が多い
    - システムを変更するたびに別の翻訳結果

# BLEU [Papineni et al. 2002]

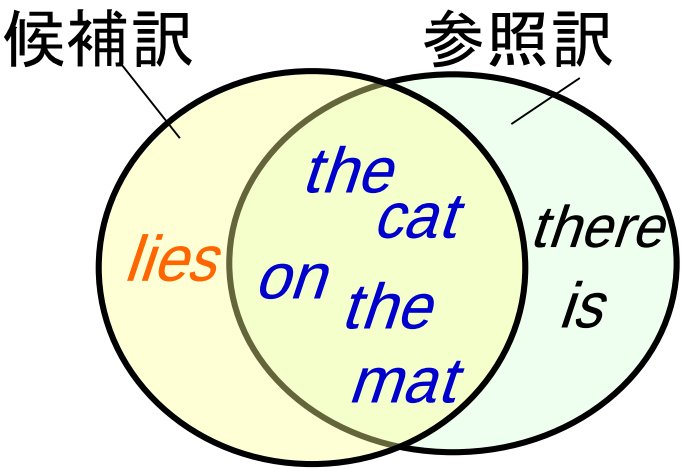
機械翻訳のための自動評価指標

- 精度ベース
- 機械翻訳の出力（候補訳）と
  人間による翻訳（参照訳）を比較

候補訳: *The cat lies on the mat.*

参照訳

1. *The cat is one the mat.*
2. *There is the cat on the mat*



# BLEU [Papineni et al. 2002]

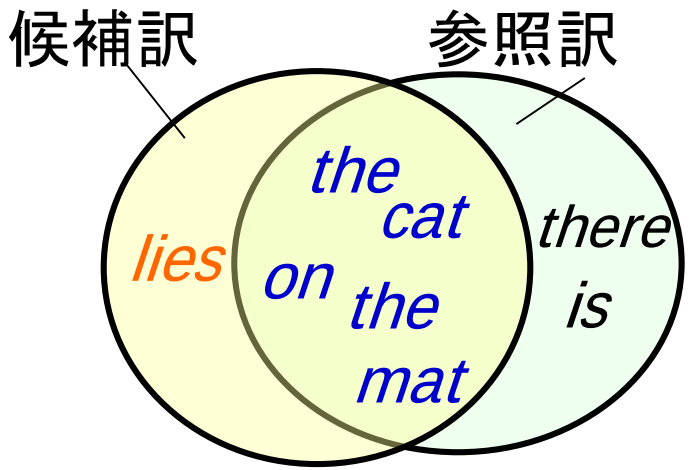
$$p = \frac{\text{候補訳と参照訳の両方にある語数}}{\text{候補訳の語数}}$$

$$p = \frac{5}{6}$$

候補訳: *The cat lies on the mat.*

参照訳

1. *The cat is one the mat.*
2. *There is the cat on the mat*



# 不適切な候補訳への対応

$$p = \frac{\text{候補訳と参照訳の両方にある語数}}{\text{候補訳の語数}}$$

$$p = \frac{6}{6}$$

候補訳: *The the the the the the.*

参照訳

1. *The cat is one the mat.*
2. *There is the cat on the mat*

分子を修正

# 不適切な候補訳への対応

候補訳と参照訳に共起した回数

$$p = \frac{\sum_{S \in Candidates} \sum_{w \in S} Count_{clip}(w)}{\sum_{S \in Candidates} \sum_{w \in S} Count(w)}$$

候補訳に出現した回数

候補訳: *The the the the the the.*

$$p = \frac{2}{6}$$

分子を修正

参照訳

1. *The cat is one the mat.*
2. *There is the cat on the mat*

$Count_{clip}(w)$ =max（候補訳中の出現数, 参照訳1文中での最大出現数）

# 正確性と流暢性

- 正確性 (adequacy)
  - 訳文が原文の内容をどれだけ保持しているか
  - 単語の訳の正確さで評価
- 流暢性 (fluency)
  - 訳文の文としての自然さ、滑らかさ
  - 単語の並びで評価

単語からn-gramへ

# 短文へのペナルティ

- 精度ベースのため、翻訳不可能な部分を無視すると、評価値が上がる
- 短文へのペナルティの導入

  $c$: 候補訳の長さ

  $r$: 参照訳の長さ

$$\mathrm{BP} = \begin{cases} 1, & c > r \\ e^{(1-r/c)}, & c \leq r \end{cases}$$

# BLEU

$$p_n = \frac{\sum_{S \in Candidates} \sum_{n\text{-}gram \in S} Count_{clip}(n\text{-}gram)}{\sum_{S \in Candidates} \sum_{n\text{-}gram \in S} Count(n\text{-}gram)}$$

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

N = 4

$w_n = 1/N$

- 小さな $n$: 正確性を評価
- 大きな $n$: 流暢性を評価

# 適用上の注意

- 異なる手法間の比較には適さない
  - ルールベースと統計ベースの比較には不適
  - 同じシステムのパラメータ改良に使用
- 文書単位で比較する
  - 文単位では長いn-gramの値が0になる
- 複数の参照訳を前提にする
  - 現実には参照訳が一つのことが多い

# その他の評価手法（1）

- WER (Word Error Rate)
  - 参照訳との編集距離を考慮

$$\mathrm{WER} = \frac{\text{置換数} + \text{挿入数} + \text{削除数}}{\text{参照訳の語数}}$$

- PER (Position independent WER)
  - 語順を無視した WER
  - 分子は厳密には距離ではない

上記の二つの手法は、1から引いた値の場合も

# その他の評価手法(2)

- NIST
  - BLEUの改良版
  - 分子に下記の値を用いる

$$\mathrm{Info}(w_1 \cdots w_n) = \frac{\text{候補訳中の } w_1 \cdots w_{n-1} \text{の数}}{\text{候補訳中の } w_1 \cdots w_n \text{の数}}$$

- ROUGE
  - 精度ベースではなく再現率ベース
  - 自動要約の評価指標

# その他の評価手法(3)

- METEOR
  - 精度と再現率の調和平均を利用
  - 同義語の情報を利用

- RIBES
  - 順位相関係数を用いる
  - 語順を考慮
  - 日英など語順が異なる言語間の翻訳評価用

# 評価指標の評価

- 人間の評価との比較
- 相関係数を計算

# 相関係数

- 二つの確率変数の間の相関の度合い
- 通常 1から -1の値をとる
  - 1に近い:　正の相関がある
  - -1に近い:　負の相関がある
  - 0に近い:　相関が弱い

- 相関がある⇒因果関係がある、ではない
  - 「アイスクリームの消費量」と「溺死者数」の間には強い相関があるが因果関係はない

# ピアソンの積率相関係数

- パラメトリックな指標
- 正規分布を仮定
- 2組の数値からなるデータ列

$$\{(x_i, y_i)\} \ (i = 1,2,\cdots,n)$$

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

# スピアマンの順位相関係数

- ノンパラメトリックな指標
- 順位が分かれば計算できる
- 分布の仮定はない

$$1 - \frac{6\sum d^2}{n^3 - n}$$

ただし、$d$ は対応するモノの順位の差

# 人手評価と自動評価指標

正確性に関する評価の相関係数(RBMTを除く)

| | | Spearman | Pearson |
|---|---|---|---|
| JE | RIBES | 0.88 | 0.96 |
| | BLEU | 0.69 | 0.83 |
| | NIST | 0.65 | 0.82 |
| EJ | RIBES | 0.93 | 0.92 |
| | BLEU | 0.76 | 0.84 |
| | NIST | 0.59 | 0.73 |

I. Goto, et al.: Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop, Proc. of the 10th NTCIR Conf. (2013)

# まとめ

- 各種の評価指標
  - 正解率･精度･再現率
- 有意差検定
- 翻訳手法の評価方法