

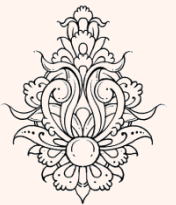
یادگیری ماشین



دانشگاه شهید بهشتی
پژوهشکده‌ی فضای مجازی
زمستان ۱۴۰۱
احمد محمودی ازناوه

فهرست مطالب

- شیوه‌های مختلف ارزیابی دسته‌بند
- روش‌های تقسیم مجموعه‌ی داده‌ها
- کارایی
- معیارهای ارزیابی

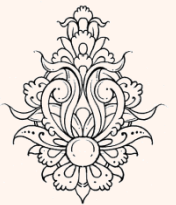


پیش‌گفتار

- با توجه به الگوریتم‌های دسته‌بندی مختلف و تأثیر هاپرپارامترها بر روی عملکرد یک روش دسته‌بندی مقایسه و انتخاب بهترین الگوریتم اجتناب‌ناپذیر است.
- تاکنون با توابع خطای متفاوتی نظیر

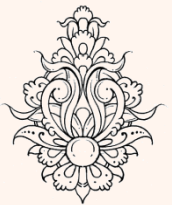
- cross-entropy
- mse
- Hing loss

آشنا شدیم، هر چند توابع خطا هم در بسیاری از موارد برای سنجش کارایی قابل استفاده هستند، اما محدود به چنین معیارهایی نیستیم. در بحث توابع خطا معمولاً بسته به نوع بهینه‌سازی) مشتق‌پذیری مورد توجه است. حال آن که معیارهای ارزیابی مطلوبی هستند که الزاماً مشتق‌پذیر هم نیستند.



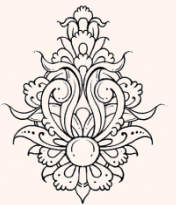
ارزیابی کارایی

- بررسی نتیجه‌ی دسته‌بند بر روی داده‌های آموزشی پذیرفتنی نیست، بلکه باید قدرت تصمیم‌پذیری روش را مورد ارزیابی قرار داد.
- معیار کارایی یک الگوریتم، خطای دسته‌بندی بر روی داده‌های آزمون (test) است.
- داده‌هایی که در مرحله‌ی آموزش مورد استفاده قرار نگرفته‌اند.
- حتی با جدا کردن داده‌های آموزش و آزمون یک بار بررسی کفایت نمی‌کند.
- ممکن است مجموعه‌ی داده‌ها کم بوده و دارای داده‌هایی خاص (نویز و داده‌های دورافتاده) باشند که بر نتیجه‌ی کلی اثر گذار هستند.
- الگوریتم آموزش به مقادیر اولیه وابسته باشد.



ارزیابی کارایی

- برای ایجاد یک دسته‌بند، یک الگوریتم دسته‌بندی و یک مجموعه داده‌ی آموزشی مورد استفاده قرار می‌گیرد.
- برای کاهش اثر عوامل تصادفی (داده‌های آموزشی، وزن‌های اولیه) یک الگوریتم برای ایجاد دسته‌بندهای متفاوت استفاده می‌شود.
- ارزیابی‌ها بر اساس توزیع خطا دسته‌بندهای متفاوت صورت می‌پذیرد.



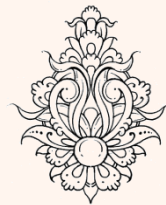
ارزیابی کارایی (ادامه...)

- باید توجه داشت که اعتبار این ارزیابی محدود به مجموعه داده‌ی مورد استفاده و کاربرد می‌شود و به معنای مقایسه‌ی کلی نمی‌باشد
- «هیچ الگوریتم بهینه‌ای برای تمام حالات وجود ندارد.»
- معمولاً داده به سه قسمت تقسیم می‌شود.

– یک قسمت برای آزمون

– دو قسمت برای آموزش و اعتبارسنجی

- شبکه عصبی: داده آموزشی برای تنظیم وزن‌ها، داده‌های اعتبارسنجی برای تنظیم واحدهای مخفی و نرخ یادگیری و داده‌های آزمون برای ارزیابی نهایی
- Knn: برای تنظیم k از داده‌های اعتبارسنجی استفاده می‌شود.

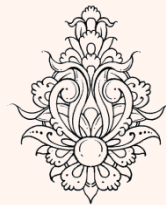


The NFL stated that within certain constraints, over the space of all possible problems, every optimization technique will perform as well as every other one on average (including Random Search)

Cross Validation

- در تقسیم داده‌ها به دو قسمت باید توجه داشت که توزیع داده‌ها حفظ شود، بدین ترتیب دانش پیشین در مورد یک کلاس دستخوش تغییر نخواهد شد.

stratification



K-fold cross validation

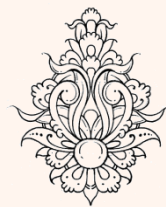
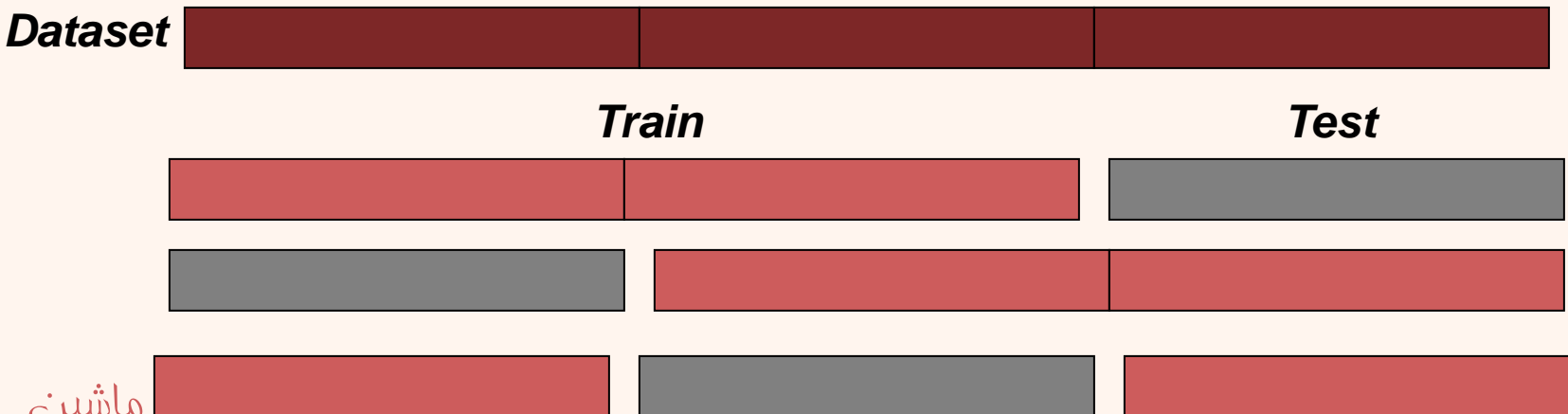
- در این شیوه مجموعه داده‌ها به k قسمت مساوی تقسیم می‌شود. هر بار یک قسمت برای اعتبارسنجی و مابقی به عنوان مجموعه آموزشی به کار می‌رود.

$$\mathcal{V}_1 = \mathcal{X}_1 \quad \mathcal{T}_1 = \mathcal{X}_2 \cup \mathcal{X}_3 \cup \dots \cup \mathcal{X}_k$$

$$\mathcal{V}_2 = \mathcal{X}_2 \quad \mathcal{T}_2 = \mathcal{X}_1 \cup \mathcal{X}_3 \cup \dots \cup \mathcal{X}_k$$

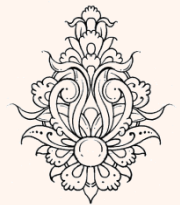
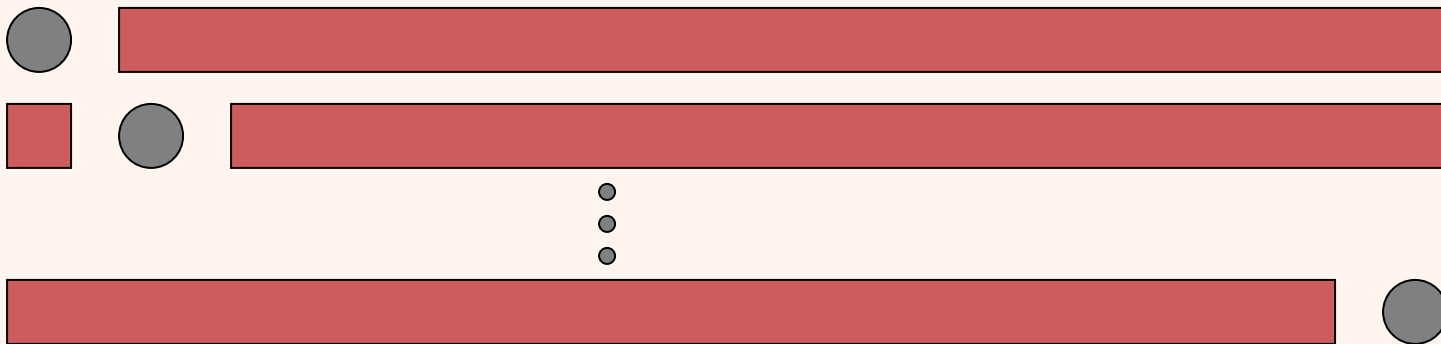
⋮

$$\mathcal{V}_k = \mathcal{X}_k \quad \mathcal{T}_k = \mathcal{X}_1 \cup \mathcal{X}_2 \cup \dots \cup \mathcal{X}_{k-1}$$



Leave-one-out (N-fold cross validation)

- یک حالت خاص k-fold است ($k=N$).
- این شیوه معمولاً در مواردی مورد استفاده قرار می‌گیرد که تهیهی داده‌ی برچسب خورده دشوار باشد. مانند تشخیص‌های پزشکی



5×2 Cross-Validation

$$\mathcal{T}_1 = \mathcal{X}_1^{(1)} \quad \mathcal{V}_1 = \mathcal{X}_1^{(2)}$$

$$\mathcal{T}_2 = \mathcal{X}_1^{(2)} \quad \mathcal{V}_2 = \mathcal{X}_1^{(1)}$$

$$\mathcal{T}_3 = \mathcal{X}_2^{(1)} \quad \mathcal{V}_3 = \mathcal{X}_2^{(2)}$$

$$\mathcal{T}_4 = \mathcal{X}_2^{(2)} \quad \mathcal{V}_4 = \mathcal{X}_2^{(1)}$$

⋮

$$\mathcal{T}_9 = \mathcal{X}_5^{(1)} \quad \mathcal{V}_9 = \mathcal{X}_5^{(2)}$$

$$\mathcal{T}_{10} = \mathcal{X}_5^{(2)} \quad \mathcal{V}_{10} = \mathcal{X}_5^{(1)}$$



5 times 2 fold cross-validation (Dietterich, 1998)

- در این شیوه N نمونه با جای‌گذاری انتخاب می‌شود.

- احتمال انتخاب یک نمونه $1/N$ است.

- احتمال این که یک نمونه N بار انتخاب نشود:

$$\left(1 - \frac{1}{N}\right)^N \approx e^{-1} = 0.368$$

- در نتیجه می‌توان گفت هر بار مجموعه‌ی آموزشی تنها ۳۶ درصد داده‌ها را در بر می‌گیرد.



ارزیابی شیوه‌های مختلف دسته‌بندی

Performance

• کارایی:

– پیش‌بینی درست برچسب کلاس

Time and Space complexity

• پیچیدگی زمانی و مکانی:

– زمان (حافظه) مورد نیاز برای آموزش

– زمان (حافظه) مورد نیاز برای دسته‌بندی

Robustness

• مقاومت:

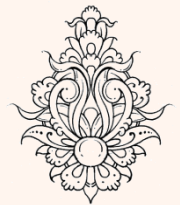
– مقاومت در برابر نویز (برچسب‌های اشتباه)، عدم وجود برخی مؤلفه‌ها

Interpretability

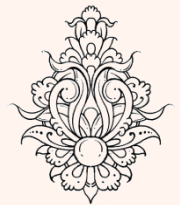
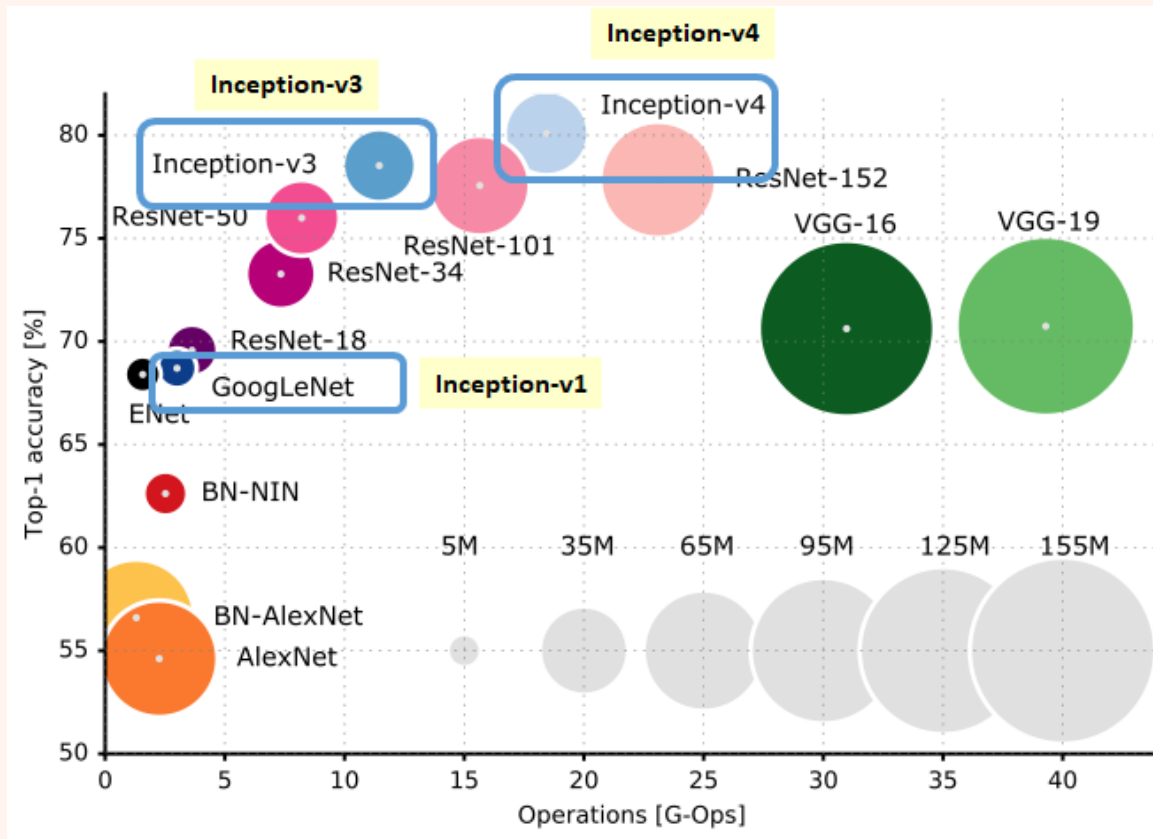
• تفسیرپذیری:

– امکان استخراج دانش

• معیارهای دیگری نظیر ریسک دسته‌بندی و سادگی نیز دارای اهمیت می‌باشند.



کارایی در برابر حجم محاسبات



<https://towardsdatascience.com/review-inception-v4-evolved-from-googlenet-merged-with-resnet-idea-image-classification-5e8c339d18bc>

5e8c339d18bc

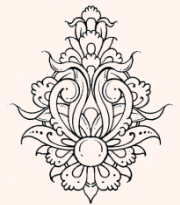
معیارهای ارزیابی

Confusion Matrix(CM)/error matrix

- **ماتریس خطا:** در این ماتریس مؤلفه‌ی $CM_{i,j}$ بیانگر این است که چند عنصر از کلاس i به عنوان عضو j از دسته‌ی j برچسب خورده است.

– قطر اصلی دسته‌بندی درست را نشان می‌دهد.

| | Ball | Car | Dri | Feed | Left | right | Pet | Shake | Sniff | Walk |
|-----------|------|-------|------|------|------|-------|-------|-------|-------|-------|
| Ball | 4.93 | 0 | 0 | 0.03 | 0.01 | 0 | 0 | 1.08 | 0.77 | 0.18 |
| Car | 0 | 12.62 | 0.06 | 0.03 | 0.04 | 0 | 0.07 | 0 | 0.18 | 0 |
| Drink | 0 | 0.45 | 3.26 | 0.35 | 0.02 | 0.01 | 0.17 | 0 | 0.72 | 0.02 |
| Feed | 0.2 | 0.24 | 0.46 | 7.61 | 0.95 | 0.3 | 1.84 | 0.2 | 0.35 | 0.85 |
| LookLeft | 0.51 | 0.94 | 0 | 1.01 | 3.76 | 2.43 | 0.21 | 0.36 | 0.05 | 1.73 |
| LookRight | 0 | 1.04 | 0 | 0.33 | 0.72 | 4.66 | 0.38 | 0 | 0.43 | 1.44 |
| Pet | 0 | 0.57 | 0 | 0.65 | 0.17 | 0.03 | 11.47 | 0.01 | 0.1 | 0 |
| Shake | 0.09 | 0 | 0 | 0.08 | 0.01 | 0 | 0.23 | 8.59 | 0 | 0 |
| Sniff | 0.04 | 0.06 | 0.01 | 0.07 | 0.09 | 0.09 | 0.04 | 0 | 13.14 | 0.46 |
| Walk | 0.03 | 0 | 0.02 | 0.05 | 0.45 | 0.11 | 0 | 0 | 0.85 | 11.49 |



معیارهای ارزیابی (ادامه...)

- در دسته‌بندی دوکلاسی این ماتریس چهار عنصر دارد:

True Positive

– مثبت صحیح

- شخص بیمار، به درستی بیمار تشخیص داده شود.

False Positive

– مثبت کاذب:

- شخص سالم، به اشتباه بیمار تشخیص داده شود.

True Negative

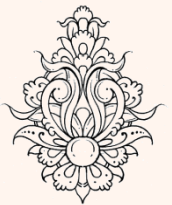
– منفی صحیح

- شخص سالم، به درستی سالم تشخیص داده شود.

False Negative

– منفی کاذب

- شخص بیمار، به اشتباه سالم تشخیص داده شود.



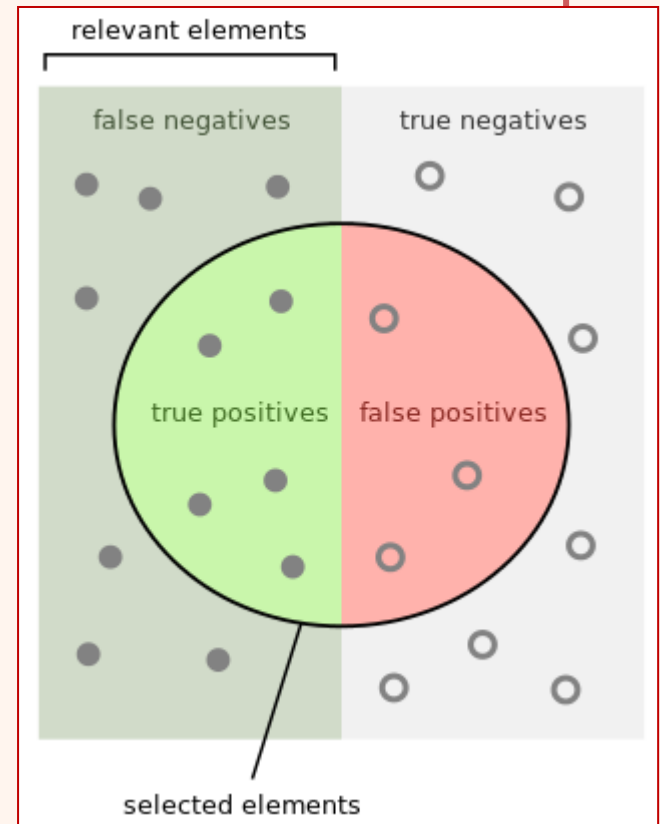
معیارهای ارزیابی (ادامه...)

| Actual Class | Predicted Class | |
|--------------|-----------------|-------|
| | True | False |
| True | TP | FN |
| False | FP | TN |

Error Type II (points to FN)

Error Type I (points to FP)

P
 N



wiki

سازمان
سازمان
سازمان

معیارهای ارزیابی (ادامه...)

- درستی (accuracy): نسبت نمونه‌های که برچسب درست خورده‌اند به کل نمونه‌ها

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{P + N}$$

- نرخ خطا (error rate): $error_rate = 1 - accuracy = \frac{FP + FN}{P + N}$

True Positive Rate (TPR)

- حساسیت (Sensitivity)، بازیابی (recall):

$$sensitivity = \frac{TP}{TP + FN} = \frac{TP}{P}$$

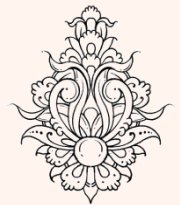
- تشخیص (ویژگی) (specificity):

$$specificity = \frac{TN}{TN + FP} = \frac{TN}{N}$$

True Negative Rate (TNR)

حساسیت به معنای احتمال درست تشخیص بیماری است

تشخیص به معنای احتمال درست تشخیص سالم بودن است

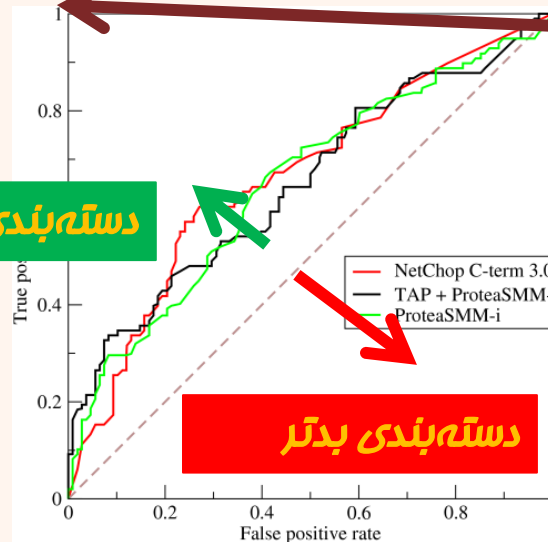
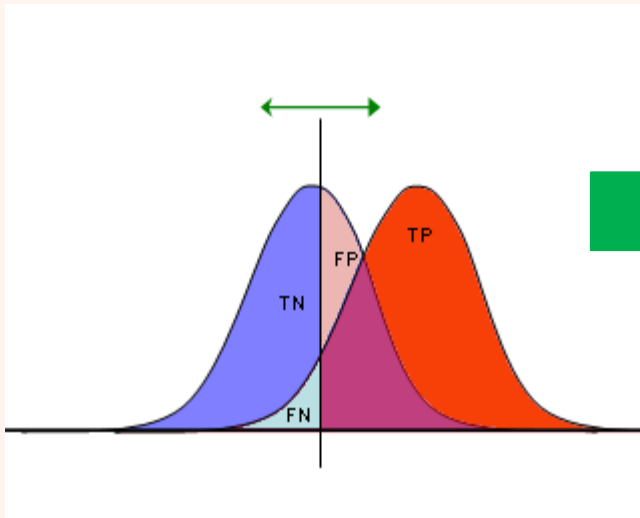


منحنی مشخصه عملکرد سیستم

Receiver Operating Characteristics (ROC)

• این منحنی رابطه‌ی بین TPR و FPR را نشان می‌دهد، زمانی که حدآستانه‌ی جداسازی تغییر می‌کند.

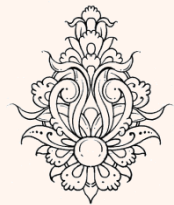
- در واقع این منحنی ابزاری است که می‌تواند برای انتخاب حدآستانه‌ی بهینه به کار رود.



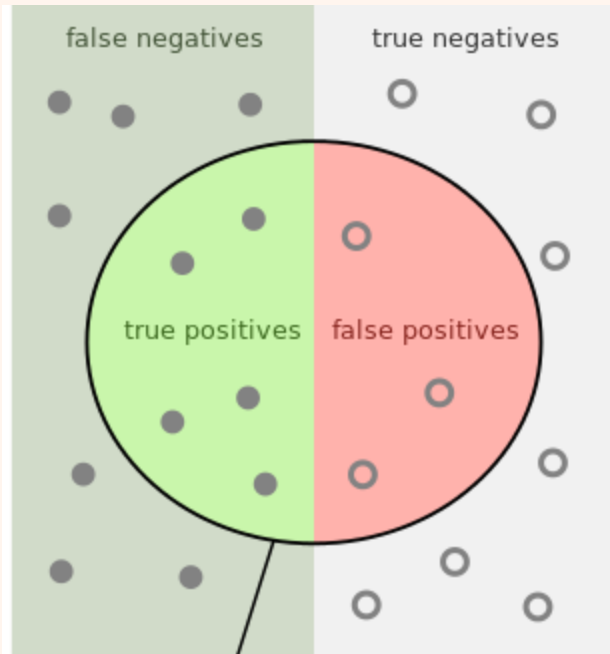
دسته بندی ایده آل

دسته بندی بهتر

دسته بندی بدتر



معیارهای ارزیابی (ادامه...)



selected elements

How many selected items are relevant?

Precision =



How many relevant items are selected?

Recall =



wiki

• بازایی (recall):

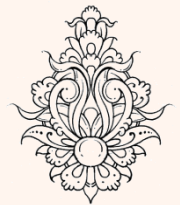
$$recall = \frac{TP}{TP + FN} = \frac{TP}{P}$$

• دقت (precision):

$$precision = \frac{TP}{TP + FP}$$

• معیار F1 (F₁ Score):

$$F_1 = \frac{2TP}{2TP + FP + FN}$$



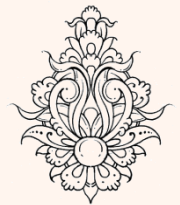
معیار F_1 (F₁ Score):

- این معیار در واقع میانگین هارمونیک دقت و بازیابی است:

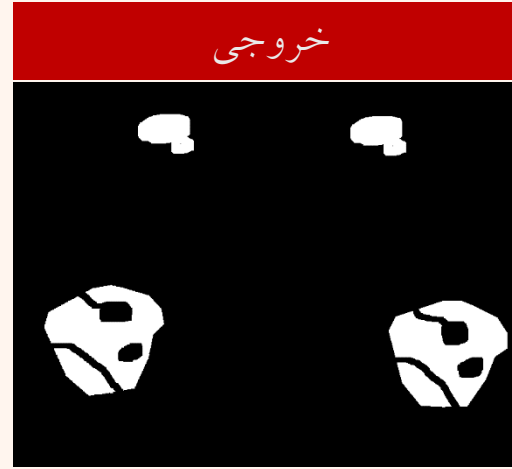
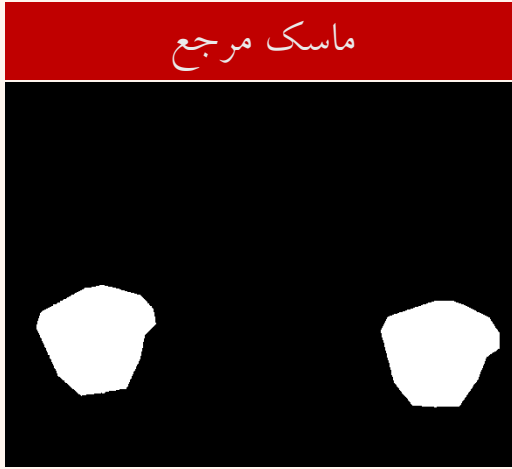
- در بین میانگین‌های فیثاغورثی کم‌ترین مقدار را دارد.

$$F_1 = \frac{2 \textit{precision} \times \textit{recall}}{\textit{precision} + \textit{recall}} = \frac{2}{\frac{1}{\textit{precision}} + \frac{1}{\textit{recall}}}$$

- با توجه به تغییر حد آستانه دو مقدار دقت و بازیابی تغییر می‌کنند، معمولاً از این معیار برای مقایسه‌ی دسته‌بندها استفاده می‌شود.



مثال



گزارش میانگین هر معیار

$$\text{Recall} = \frac{|\{\text{Forged Pixels}\} \cap \{\text{Detected Pixels}\}|}{|\{\text{Forged Pixels}\}|}$$

$$\text{Precision} = \frac{|\{\text{Forged Pixels}\} \cap \{\text{Detected Pixels}\}|}{|\{\text{Detected Pixels}\}|}$$

$$F_1 = 2 \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

