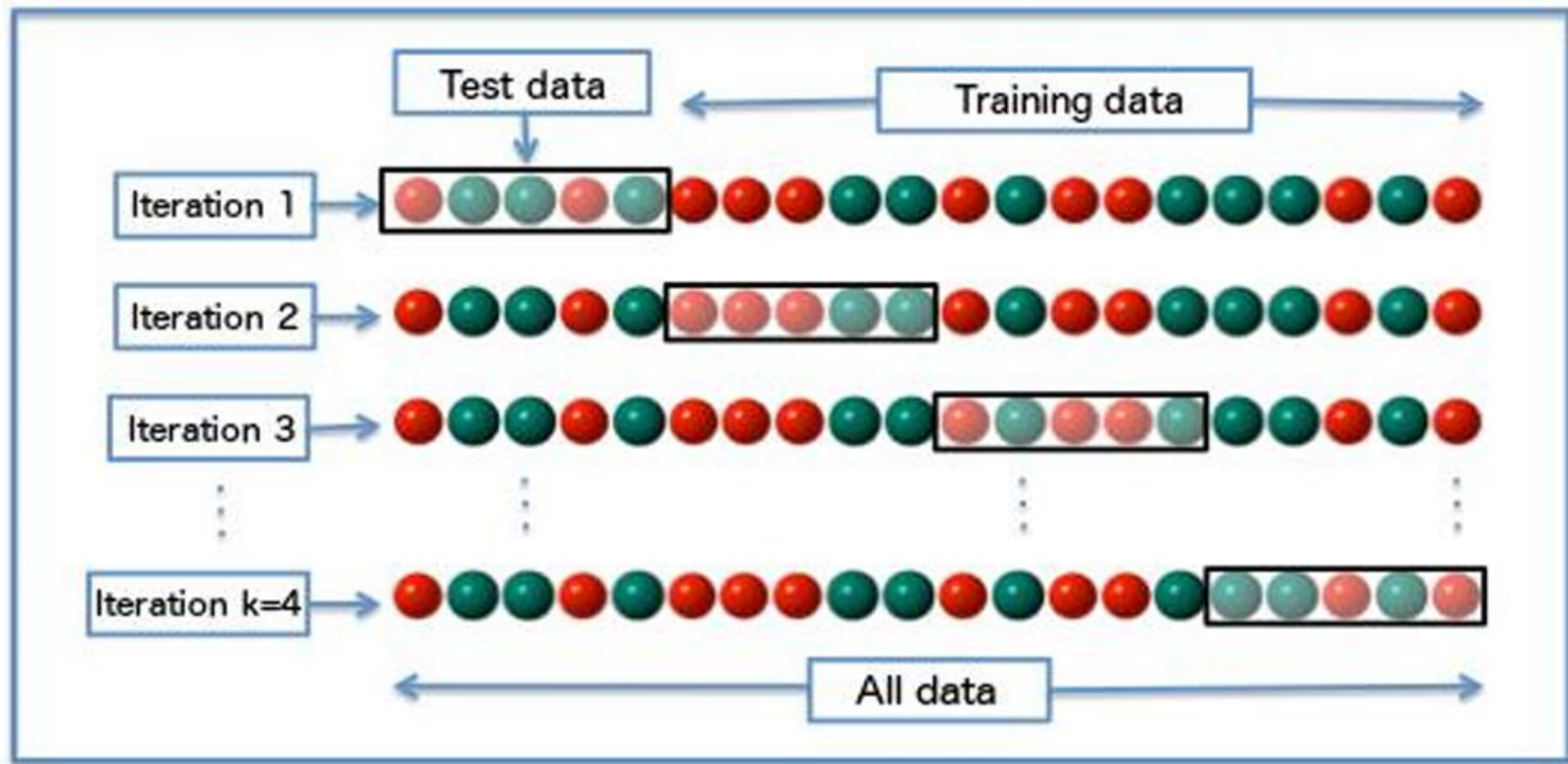


K-Folds الطبقات العديدة (9)



في البداية علينا ان نتعرف علي أنواع الـ Cross-validation او باختصارها : CV :


● الأنواع الغير شاملة

Non-Exhaustive CV (التي لا يتم عمل التدريب علي كامل العينة)

- الطبقات العديدة
 - هي درس اليوم
- اسلوب التحمل
 - الطريقة العادية في تقسيم البيانات لقسم تدريب و قسم اختبار بنسب معينة
- العينات العشوائية المتكررة
 - يتم تقسيم العينة لعدد من الأجزاء , وكل جزء يتم تقسيمه هو نفسه الي عينة تدريب و عينة اختبار , و من ثم تكون النتيجة النهائية هي متوسط نتائج عينات الاختبار , وتسمى طريقة مونت كارلو

✓ k-fold cross-validation

✓ Holdout method



● الأنواع الشاملة

Exhaustive CV (التي يتم عمل التدريب و الاختبار علي كل العينة)

- الأجزاء المتروكة
 - يتم اختيار عدد من العينة لتكون للاختبار و الباقي للتدريب , ثم التكرار مرة اخري
- الجزء الواحد المتروك
 - نفس الفكرة السابقة , لكن يكون العدد يساوي 1

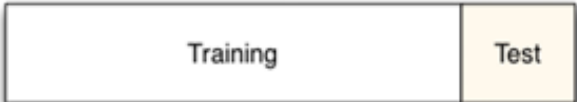
(LpO CV) Leave-p-out cross-validation

(LOOCV) Leave-one-out cross-validation

إذن ما هي فكرة الـ K-fold

هي أحد الطرق المستخدمة لتقسيم البيانات , بدلا من عمل training & test data ان يتم تناول بيانات التدريب , و تقسيمها عدد من اجزاء (او تطبيقات folds)

iter 1



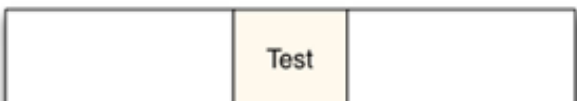
فلو كان لدينا 5000 طالب , يكون القسم الأول من 1:500 , الثاني من 501 إلي 1000 , والعاشر من 4501 إلي 5000 , وكل قسم فيهم يسمى fold

iter 2



ثم نقوم في المحاولة الأولى بجعل القسم الأول (طالب 1 إلي 500) هي بيانات الاختبار في حين باقي الاقسام التسعة هي بيانات التدريب . .

iter 3



ثم تكرر الأمر في المحاولة الثانية , لكن ان يكون القسم الثاني (501-1000) هو الاختبار في حين القسم الأول و من الثاني للعاشر هي التدريب .

iter 4



و يتم تكرار الأمر عشر مرات

iter 5



و في النهاية نقوم بحساب متوسط الخطأ الناتج من كل فولد منهم , حتي نصل للخطأ المتوسط لهم جميعا

و هنا يكون مثال لتوزيع البيانات

Iteration	Training set observations	Testing set observations
1	[5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24]	[0 1 2 3 4]
2	[0 1 2 3 4 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24]	[5 6 7 8 9]
3	[0 1 2 3 4 5 6 7 8 9 15 16 17 18 19 20 21 22 23 24]	[10 11 12 13 14]
4	[0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 20 21 22 23 24]	[15 16 17 18 19]
5	[0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19]	[20 21 22 23 24]

مع اعتبار عدد من الملحوظات الهامة :

- الرقم المفضل للاستخدام لك k هو 10
- اذا ما ستقوم بعمل هذا في التصنيف classification فجيب ان تراعي ان يكون كل قسم fold ان يكون به نسب متوازيه من القسمين
فإذا كان في البيانات لديك نسبة 15 % مرضي و 85 % غير مرضي , و لديك 10 الاف مريض , أي أن لديك فعليا 1500 مريض , و 8500 سليم
- فحينما تقوم بعمل التقسيمات folds راعي ان يكون في كل قسم فيهم نفس النسبة , أي أنه في القسم الأول الذي به الف شخص , يكون فيه 150 مريض و 850 سليم , وهكذا
- و مكتبة sklearn تقوم بهذا بال default , وهذا الأمر يسمى stratified sampling
- من الممكن تكرار عملية الـ k-fold عدد من المرات بعدد مختلف من الفولدر , لتحديد ايهم افضل
- أحيانا يتم اقتطاع جزء من العينة, والتي لا يتم تضمينها ضمن العينة المقسمة في الـ k-fold , فلو كانت العينة كلها 10 الاف شخص , يتم اقتطاع الف منها جانبا , ثم القيام بتقسيم الالف التسعة كما ذكرنا
- و ميزتها , اننا نقوم فيها بعمل اختبار نهائي للخوارزم بعد الانتهاء منه , او حتي بعد اختيار عدد مناسب من المعاملات العليا , ونحن نضمن أن هذه بيانات لم يتم مسها بعد , فالمصادقية تكون اعلي
- غالبا يتم استخدام stratifiedkfold و التي تجعل الفارق الاساسي بينها و بين kfolds العادية , انها تقوم بتقسيم الطبقات بحيث يكون لكل قسم فيهم كميات متوازنة بين جميع الاصناف