**CROS VALIDATION**

->The better for validation data

* Introduction

Pembelajaran mesin adalah proses berulang.

Anda akan menghadapi pilihan tentang variabel prediktif apa yang akan digunakan, jenis model apa yang digunakan, argumen apa yang akan diberikan ke model tersebut, dll. Sejauh ini, Anda telah membuat pilihan ini berdasarkan data dengan mengukur kualitas model dengan validasi ( atau ketidaksepakatan) ditetapkan.

Tetapi ada beberapa kelemahan dari pendekatan ini. Untuk melihat ini, bayangkan Anda memiliki kumpulan data dengan 5000 baris. Anda biasanya akan menyimpan sekitar 20% data sebagai set data validasi, atau 1000 baris. Tapi ini menyisakan beberapa **peluang acak** dalam menentukan skor model. Artinya, model mungkin berhasil pada satu set 1000 baris, bahkan jika itu tidak akurat pada 1000 baris yang berbeda.

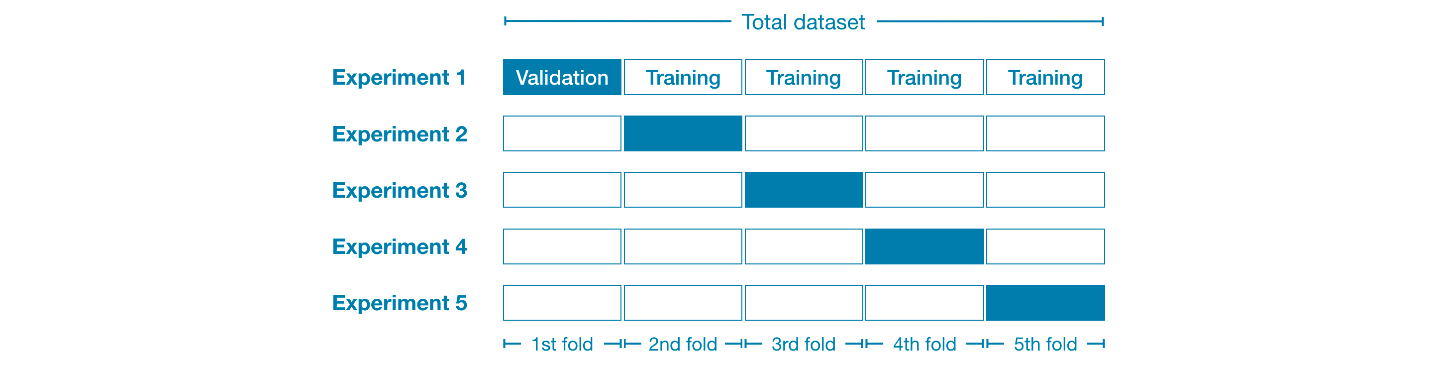
Secara ekstrem, Anda dapat membayangkan hanya memiliki 1 baris data dalam set validasi. Jika Anda membandingkan model alternatif, model mana yang membuat prediksi terbaik pada satu titik data sebagian besar adalah masalah keberuntungan!

Secara umum, **semakin besar set validasi**, semakin sedikit keacakan (alias "noise") dalam ukuran kualitas model kami, dan akan semakin andal. Sayangnya, kami hanya bisa **mendapatkan set validasi besar** dengan **menghapus baris dari data pelatihan** kami, dan **set data pelatihan yang lebih kecil** berarti model yang lebih buruk!

* What is cross validation

Pada cross validation, kami menjalankan proses pemodelan kami pada subset data yang berbeda untuk mendapatkan beberapa ukuran kualitas model.

Misalnya, kita bisa mulai dengan membagi data menjadi 5 bagian, masing-masing 20% dari kumpulan data lengkap. Dalam hal ini, kami mengatakan bahwa kami telah memecah data menjadi 5 "lipatan/**folds**".



Splitting train and validation state dipisah dengan cv, yg mana di atas adalah cv=5

Kemudian, kami menjalankan satu eksperimen untuk setiap lipatan:

1. In experiment 1, kami menggunakan lipatan pertama sebagai set validasi (atau ketidaksepakatan/Holdout) dan yang lainnya sebagai data pelatihan. Ini memberi kita ukuran kualitas model berdasarkan set ketidaksepakatan 20%.
2. In experiment 2, kami menyimpan data dari lipatan kedua (dan menggunakan semuanya kecuali lipatan kedua untuk melatih model). Holdout set kemudian digunakan untuk mendapatkan estimasi kedua dari kualitas model.
3. Kami mengulangi proses ini, menggunakan setiap lipatan sekali sebagai set ketidaksepakatan/holdout. Menyatukan ini, 100% data digunakan sebagai ketidaksepakatan di beberapa titik, dan kami berakhir dengan ukuran kualitas model yang didasarkan pada semua baris dalam kumpulan data (bahkan jika kami tidak menggunakan semua baris secara bersamaan) .

* When should you use cross-validation?

Jadi, dengan pengorbanan ini, kapan Anda harus menggunakan setiap pendekatan?

1. **Untuk kumpulan data kecil**, di mana beban komputasi ekstra bukanlah masalah besar, Anda harus menjalankan cross validation dikarenakan data nya tidak mencukupi.
2. **Untuk kumpulan data yang lebih besar**, satu kumpulan validasi sudah cukup. Kode Anda akan berjalan lebih cepat, dan Anda mungkin **memiliki cukup data** sehingga untuk ketidaksepakatan/holdout..

Tidak ada batasan sederhana untuk apa yang merupakan kumpulan data besar vs kecil. Tetapi jika model Anda membutuhkan waktu beberapa menit atau kurang untuk berjalan, mungkin ada baiknya beralih ke validasi silang.

Atau, Anda dapat menjalankan validasi silang dan melihat apakah skor untuk setiap eksperimen tampak mendekati. Jika setiap eksperimen menghasilkan hasil yang sama, satu set validasi mungkin sudah cukup.