Outline

Quality metrics
- Supervised machine learning problems

Generalisation

Estimating quality
- Methodology

Learning and generalisation
- Generalisation
- Bias and variance

% Created 2023-10-03 Di 08:29 % Intended LaTeX compiler: pdflatex [smaller]beamer [utf8]inputenc [T1]fontenc graphicx grffile longtable wrapfig rotating [normalem]ulem amsmath textcomp amssymb capt-of hyperref tikz amsmath amssymb isomath $\{ \{\{\mathbb{E}\}\}\} \{ \{\{\mathbb{V}\}\}\} \{ \{\{\mathbb{B}\}\}\} \{ \{\{\mathbb{I}\}\}$ \#1\} $\{\{ \{\{\mathbb{P}\}\}\}\} *$ *arg max* $*$ *arg min* $*$ *sgn* $\triangleq \}\mathbb{R}\}$ $\Pr \Theta\}$ $\Pr \theta\}$ $\theta\}\Theta\}W\}w\}w \Pr_i\}$ $w\Pr_{i,j}\}$ A} $a\Pr_i\}$ $a\Pr_{i,j}\}$ x} $\Pr \beta\}$ Bernoulli} Beta} Normal} []beamerthemedefault { pdfauthor={Christos Dimitrakakis}, pdftitle={Machine Learning and Data Mining}, pdfkeywords={}, pdfsubject={}, pdfcreator={Emacs 26.3 (Org mode 9.1.9)}, pdflang={English}}

Machine Learning and Data Mining

Christos Dimitrakakis

November 16, 2023

Outline

Quality metrics
- Supervised machine learning problems

Generalisation

Estimating quality
- Methodology

Learning and generalisation
- Generalisation
- Bias and variance

Classification

The classifier as a decision rule

A decision rule Pr $\pi(a|x)$ generates a decision Pr $a \in [m]$. It is the conditional probability of Pr a given Pr x.

A note on conditional probabilities

Even though normally conditional probabilities are defined as Pr $P(A|B) = P(A \cap B)/P(B)$, the probability of the decision Pr a is undefined without a given Pr x. So it's better to think if Pr $\pi(a|x)$ as a collection of distributions on Pr a, one for each value of Pr x.

Deterministic predictions given a model Pr $P(y|x)$

Here, we pick the most likely class:

$$\pi(a|x_t) = a =_y P(y|x_t)$$

Deterministic predictions given a model Pr $P(y|x)$

Here, we randomly select a class according to our model:

Accuracy as a classification metric

The accuracy of a single decision

$$U(a_t, y_t) = a_t = y_t = \begin{cases} 1, & \text{if } a_t = y_t \\ 0, & \text{otherwise} \end{cases}$$

The accuracy on the training set

$$U(\pi, D)\frac{1}{T}\sum_{t=1}^{T}\sum_{a=1}^{m}\pi(y_t|x_t)$$

The expected accuracy of a decision rule

If $\Pr(x, y) \sim P$, the accuracy $\Pr U$ of a stochastic decision rule $\Pr \pi$ under the distribution $\Pr P$ is the probability it predicts correctly

$$U(\pi, P)\int_X dP(x)\sum_{y=1}^{m} P(y|x)\pi(y|x)$$

Regression

The regressor as a decision rule

A decision rule $\Pr \pi$ generates a decision $\Pr a \in^m$.

- For randomised rules, $\Pr \pi(a|x)$ is the conditional density of $\Pr a$ given $\Pr x$.
- For deterministic rules $\Pr \pi(x)$ is the prediction for $\Pr x$.

Mean-Squared Error on a Dataset

The mean-square error is simply the squared difference in predicted versus actual values:

$$\frac{1}{T}\sum_{t=1}^{T}[y_t - \pi(x_t)]^2$$

Expected MSE

If $\Pr(x, y) \sim P$, the expected MSE of a deterministic decision rule $\Pr \pi : X \to$ is

Training and overfitting

Training data

- $\Pr D = ((x_t, y_t) : t = 1, \ldots, T)$.
- $\Pr x_t \in X$, $\Pr y_t \in Y$.

Assumption: The data is generated i.i.d.

- $\Pr(x_t, y_t) \sim P$ for all $\Pr t$ (identical)
- $\Pr D \sim P^T$ (independent)

The optimal decision rule for $\Pr P$

$$\max_\pi U(\pi, P) = \max_\pi \int_{X \times Y} dP(x, y) \sum_a \pi(a|x) U(a, y)$$

The optimal decision rule for $\Pr D$

$$\max_\pi U(\pi, D) = \max_\pi \sum_{(x,y) \in D} \sum_a \pi(a|x) U(a, y)$$

The Train/Validation/Test methodology

Main idea

Use each piece of data once to make decisions and measure

Training set

Use to decide low-level model parameters

Validation set

Use to decide between:

- different hyperparameters (e.g. Pr K in nearest neighbours)
- model (e.g. neural networks versus kNN)

Test set

Use to measure the final quality of a model

Cross-validation (XV)

Idea

- Use XV to select hyperparameters instead of a single train/valid test.

Methodology

- Split training set Pr D in Pr k different subsets
- At iteration Pr i
- Use the i-th subset for validation
- Use all the remaining Pr $k-1$ subsets for training
- Average results on validation sets

Bootstrapping

Idea

- How to take into account variability?
- Resample the data and repeat your calculations for each resample

Boostrap samples

- Input: Data $\Pr D$, of size $\Pr T$
- For $\Pr t$ in $\Pr\{1, \ldots, T\}$

– Select $\Pr i$ uniformly in $\Pr[T]$ – Add the i-th point to $\Pr D_b$

- Return $\Pr D_b$

The wrong way to do XV for subset selection

1. Screen the predictors: find a subset of “good” predictors that show

fairly strong (univariate) correlation with the class labels

1. Using just this subset of predictors, build a multivariate classifier.
2. Use cross-validation to estimate the unknown tuning parameters and

to estimate the prediction error of the final model. Is this a correct application of cross-validation? Consider a scenario with $N = 50$ samples in two equal-sized classes, and $p = 5000$ quantitative predictors (standard Gaussian) that are independent of the class labels. The true (test) error rate of any classifier is 50%.

The right way to do XV for feature selection

1. Divide the samples into K cross-validation folds (groups) at random.
2. For each fold Pr $k = 1, 2, \ldots, K$ (a) Find a subset of “good” predictors that show fairly strong (univariate) correlation with the class labels, using all of the samples except those in fold k. (b) Using just this subset of predictors, build a multivariate classifier, using all of the samples except those in fold k. (c) Use the classifier to predict the class labels for the samples in

fold k.

Generalisation as error

Error due to mismatched objectives

The Pr π^* maximising Pr $U(\pi, P)$ is not the Pr $\hat{\pi}$ maximising Pr $U(\pi, D)$.

Lemma

If Pr $|U(\pi, P) - U(\pi, D)| \leq \epsilon$ for all Pr π then

$$U(\hat{\pi}, D) \geq U(\pi^*, P) - 2\epsilon.$$

Error due to restricted classes

- We may use a constrained Pr $\hat{\Pi} \subset \Pi$.
- Then Pr $\max_{\hat{\pi} \in \hat{\Pi}} U(\pi, P) \leq \max_{\pi \in \Pi} U(\pi, P)$.

The bias/variance trade-off

- Dataset Pr D P.
- Predictor Pr $f_D(x)$
- Target function Pr $y = f(x) + \epsilon$
- Pr$\{\epsilon = 0$ zero-mean noise with variance Pr $\sigma^2 = (\epsilon)$

MSE decomposition

$$\{[(f - f_D)^2] = (f_D) + (f_D)^2 + \sigma^2$$

Variance

How sensitive the estimator is to the data

$$(f_D) = \{[(f_D - \{(f_D))^2]$$

Bias

What is the expected deviation from the true function

$$(f_D) = \{[(f_D - f)]$$

Example: mean estimation

- Data $\Pr D = y_1, \ldots, y_T$ with $\Pr\{[y_t] = \mu$.
- Goal: estimate $\Pr \mu$ with some estimator $\Pr f_D$ to minimise
- MSE: $\Pr\{[(y - f_D)^2]$, the expected square difference between new samples our guess.

Optimal estimate

To minimise the MSE, we use $\Pr f^* = \mu$. This gives us two ideas:

Empirical mean estimator:

- $\Pr f_D = \sum_{t=1}^T x_t / T$.
- $\Pr(f_D) = \{[f_D - \mu] = 1/\sqrt{T}$
- $\Pr(f_D) = 0$.

Laplace mean estimator:

- $\Pr f_D = \sum_{t=1}^T (\lambda + x_t) / T$.
- $\Pr(f_D) = \{[f_D - \mu] = \frac{1}{1+\sqrt{T}}$
- $\Pr(f_D) = O(1/T)$.

A proof of the bias/variance trade-off

- RV's Pr $y_t \sim P$, Pr$\{[y_t] = \mu$, Pr $y_t = \mu + \epsilon_t$.
- Estimator Pr f_D, Pr $D = y_1, \ldots, y_{t-1}$.

$$
\begin{aligned}
\{[(f_D - y_t)^2] &= \{[f_D^2] - 2\{[f_D y_t] + \{[y_t^2] \\
&= [f_D] + \{[f_D]^2 - 2\{[f_D y_t] + \{[y_t^2] \\
&= [f_D] + \{[f_D]^2 - 2\{[f_D]\{[y_t] + \{[y_t^2] \\
&= [f_D] + \{[f_D]^2 - 2\{[f_D]\mu + \{[y_t^2] \\
&= [f_D] + \{[f_D]^2 - 2\{[f_D]\mu + \{[(\mu + \epsilon_t)^2] \\
&= [f_D] + \{[f_D]^2 - 2\{[f_D]\mu + \{[\mu^2 + 2\mu\epsilon_t + \epsilon_t^2] \\
&= [f_D] + \{[f_D]^2 - 2\{[f_D]\mu + \mu^2 + \sigma^2 \\
&= [f_D] + (\{[f_D] - \mu)^2 + \sigma^2 \\
&= (f_D) + (f_D)^2 + \sigma^2
\end{aligned}
$$