DAT2 - week 6

Misrab M. Faizullah-Khan

Review

- Exponential Family, GLMs
- Cross validation and overall flow

The Exponential Family: Bernoulli

$$p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$$

$$\begin{aligned} p(y;\phi) &= \phi^y(1-\phi)^{1-y} \\ &= \exp(y\log\phi + (1-y)\log(1-\phi)) \\ &= \exp\left(\left(\log\left(\frac{\phi}{1-\phi}\right)\right)y + \log(1-\phi)\right) \end{aligned}$$

What are…

$\eta = \log(\phi/(1-\phi))$.

$$\begin{aligned} T(y) &= y \\ a(\eta) &= -\log(1-\phi) \\ &= \log(1+e^{\eta}) \\ b(y) &= 1 \end{aligned}$$

The Exponential Family: Normal

$$p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$$

$$\begin{aligned} p(y;\mu) &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(y-\mu)^2\right) \\ &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y^2\right)\cdot\exp\left(\mu y - \frac{1}{2}\mu^2\right) \end{aligned}$$

What are…

$$\begin{aligned} \eta &= \mu \\ T(y) &= y \\ a(\eta) &= \mu^2/2 \\ &= \eta^2/2 \\ b(y) &= (1/\sqrt{2\pi})\exp(-y^2/2) \end{aligned}$$

Okay, okay…so who cares?

Constructing GLMs

1. Assume y | x; θ ~ ExponentialFamily(η)

2. Given x, we want to predict T(y), usually = y.
We choose h(x) = E[y|x]

3. Further assume η = θ^T.x

So we have a machinery we can crank

Constructing GLMs

Linear Regression

$$
\begin{aligned}
h_\theta(x) &= E[y|x;\theta] \\
&= \mu \\
&= \eta \\
&= \theta^T x.
\end{aligned}
$$

Logistic Classification

$$
\begin{aligned}
h_\theta(x) &= E[y|x;\theta] \\
&= \phi \\
&= 1/(1+e^{-\eta}) \\
&= 1/(1+e^{-\theta^T x})
\end{aligned}
$$

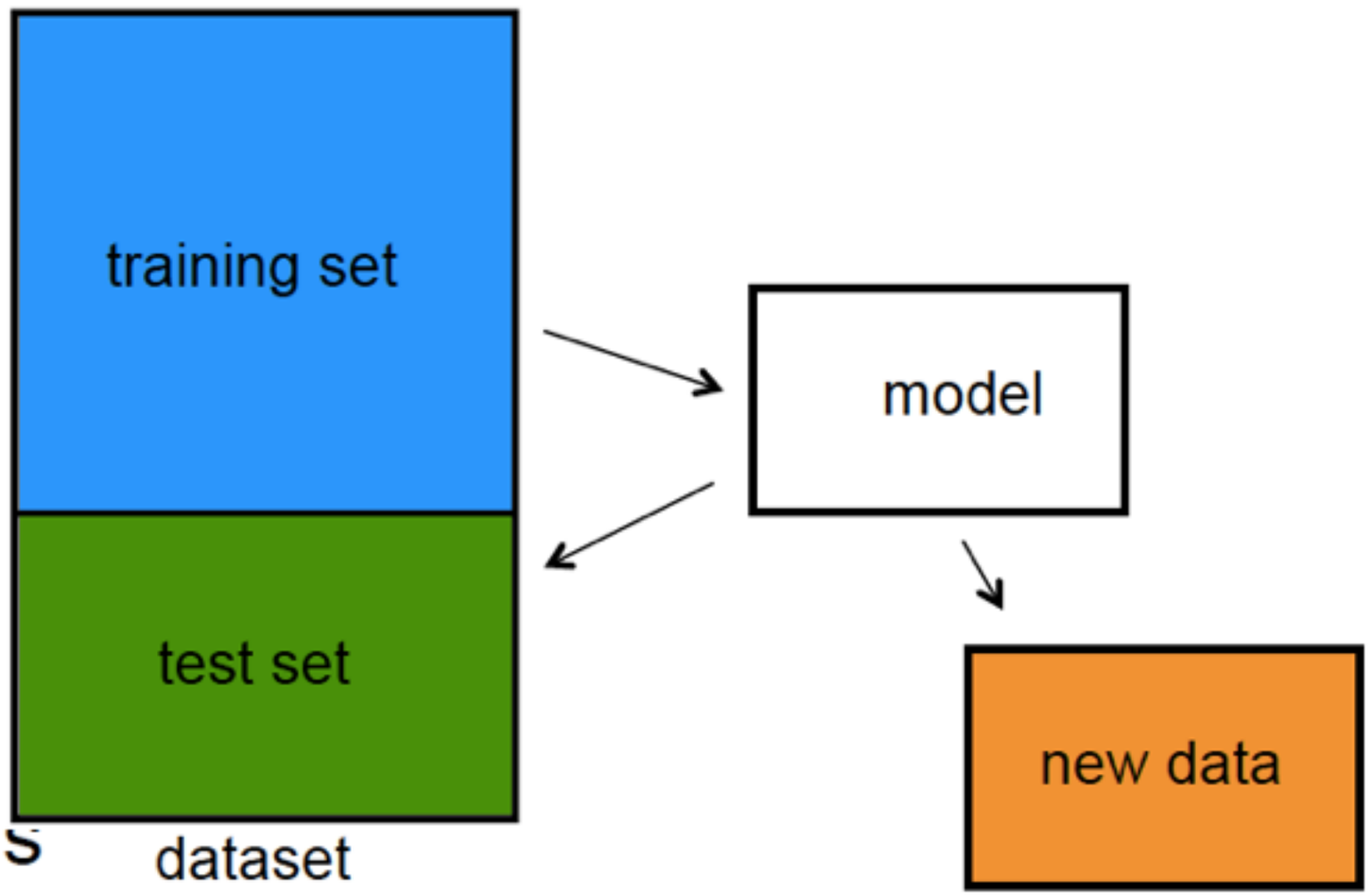
Coincidentally, this is how we get softmax regression…

Cross Validation

TEST SET APPROACH

Q: How can we make a model that generalizes well?

1) split dataset
2) train model
3) test model
4) parameter tuning
5) choose best model
6) train on **all** data
7) make predictions on new data

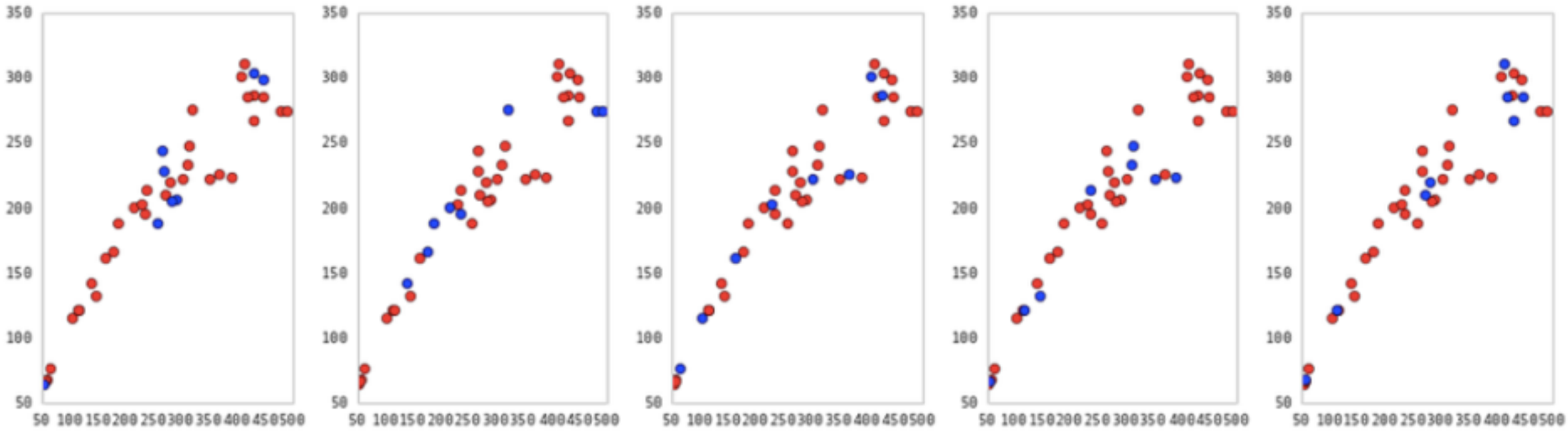


CROSS-VALIDATION

Steps for K-fold cross-validation:

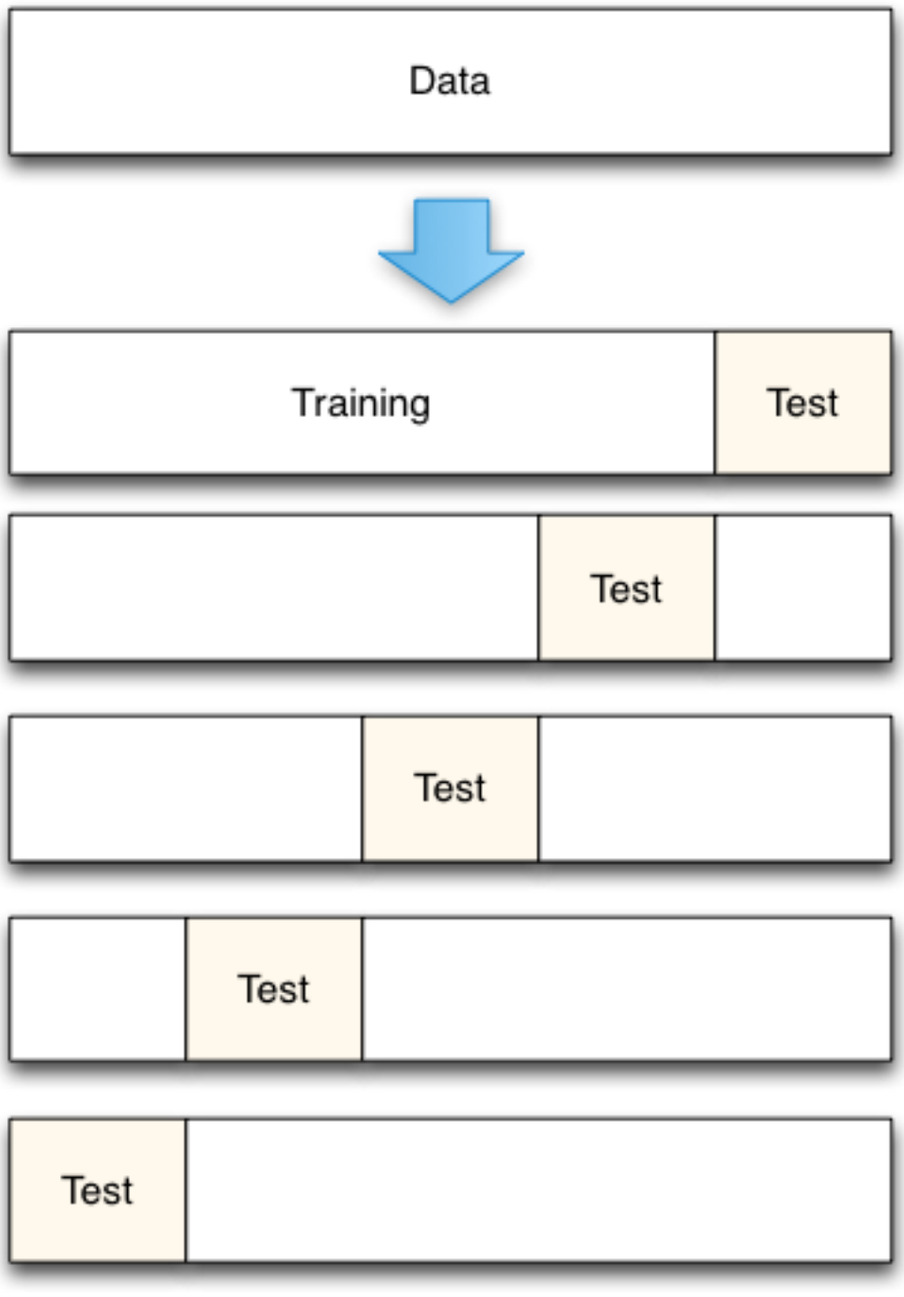
1) Randomly split the dataset into K equal partitions.
2) Use partition 1 as test set & union of other partitions as training set.
3) Calculate test set error.
4) Repeat steps 2-3 using a different partition as the test set at each iteration.
5) Take the average test set error as the estimate of OOS accuracy.

CROSS-VALIDATION



5-fold cross-validation: red = training folds, blue = test fold

Source: http://nbviewer.ipython.org/github/fonnesbeck/Bios366/blob/master/notebooks/Section6_3-Model-Selection-and-Validation.ipynb


Data
Training
Test
Test
Test
Test
Test

sklearn flow

import data, clean dataframe, visualise

↓

instantiate Model(), fit_(transform), predict

↓

cross-validation on parameters (external), features, and models