ADVANCED REGRESSION METHODS

PROS AND CONS OF MLR

…recall from the MLR lecture…

PRINCIPAL COMPONENT REGRESSION

PRINCIPAL COMPONENT REGRESSION

Let Z represent the principal components of our features

$$Z_m = \varphi_{1m}X_1 + \varphi_{2m}X_2 + \cdots + \varphi_{pm}X_p$$

Then we can fit the linear regression model

$$y = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + e_i$$

The coefficients are found with least squares method (as in MLR)

PRINCIPAL COMPONENT REGRESSION

We assume that…
…the directions in which X_1,…, X_p show the most variation are the directions that are associated with Y.
…it is not guaranteed to be true, but is often reasonable.

We estimate $M << p$ coefficients…
…and can avoid overfitting

PRINCIPAL COMPONENT REGRESSION

Increasing the number of PC-s…
…reduces bias
…but increases variance

PCR does well if first PC-s explain most of the variation

The number of PC-s is chosen by cross-validation

PARTIAL LEAST SQUARES REGRESSION

PARTIAL LEAST SQUARES

In PCR the responses do not supervise the PC-s

PLSR is an supervised alternative to PC:
…aim to find directions that help explain both the response and the features

$$Z_m = \varphi_{1m}X_1 + \varphi_{2m}X_2 + \cdots + \varphi_{pm}X_p$$

For this the coefficients φ are set to be proportional to the coefficients from the simple linear regression Y to X_j.

PARTIAL LEAST SQUARES

Compared to MLR accounts for variability in all directions.

Quite popular in classical chemometrics.

In practice is often not better then PCR.

KNN FOR CONTINUOUS TARGETS

KNN

Strictly speaking not a regression

To predict a continues target feature by a k nearest neighbor model:

$$y_{predicted} = \frac{1}{k}\sum_{i=1}^{k} y_i$$

We can use weighted knn to take into account the distance from the query

$$y_{predicted} = \frac{\sum_{i=1}^{k} \frac{1}{dist(i\ to\ pred)^2} \times t_i}{\sum_{i=1}^{k} \frac{1}{dist(i\ to\ pred)^2}}$$

KNN

Increase in k increases robustness & reduces flexibility.
…points far away from the new observation have too much weight.

Good for solving complex non-linear tasks

Variables need to be scaled

Provides NO understanding

Problem if too many variables…
… and if insignificant are in the dataset

POLYNOMINAL REGRESSION

POLYNOMINAL REGRESSION

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_d x_i^d + \epsilon_i,$$

For large enough *d* allows to produce very flexible non-linear curves
However, *d* > 3 or 4 is overly flexible

Coefficients can be estimated with least squares (not different from MLR)

STEP FUNCTION

STEP FUNCTION

MLR and polynomial regression assume a global structure

In step function we break the range into X *bins*…
…and fit a different constant in each bin
…for this we introduce cutpoints c_1, c_2, … c_K

$$
\begin{aligned}
C_0(X) &= I(X < c_1), \\
C_1(X) &= I(c_1 \leq X < c_2), \\
C_2(X) &= I(c_2 \leq X < c_3), \\
&\vdots \\
C_{K-1}(X) &= I(c_{K-1} \leq X < c_K), \\
C_K(X) &= I(c_K \leq X),
\end{aligned}
$$

STEP FUNCTION

Cutpoints are technically similar to dummy variables

$$y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \ldots + \beta_K C_K(x_i) + \epsilon_i.$$

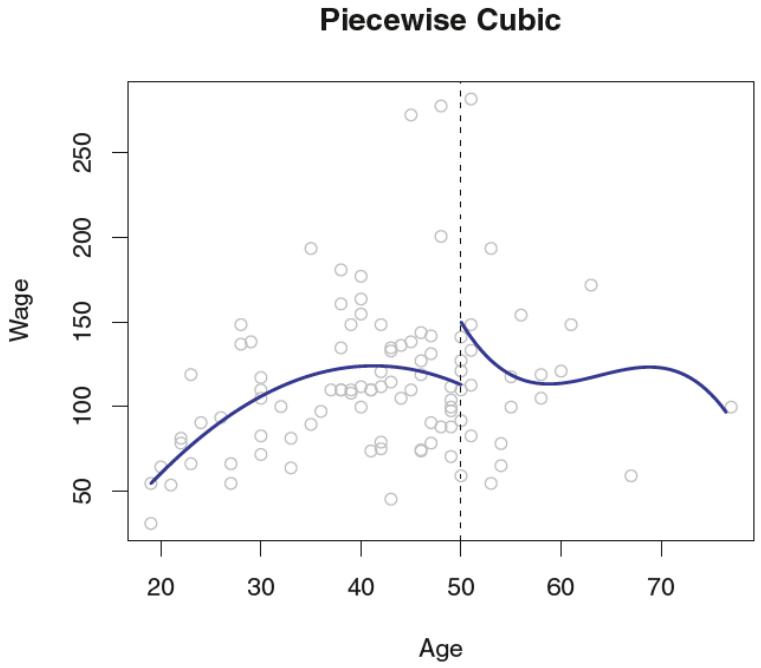
SPLINES

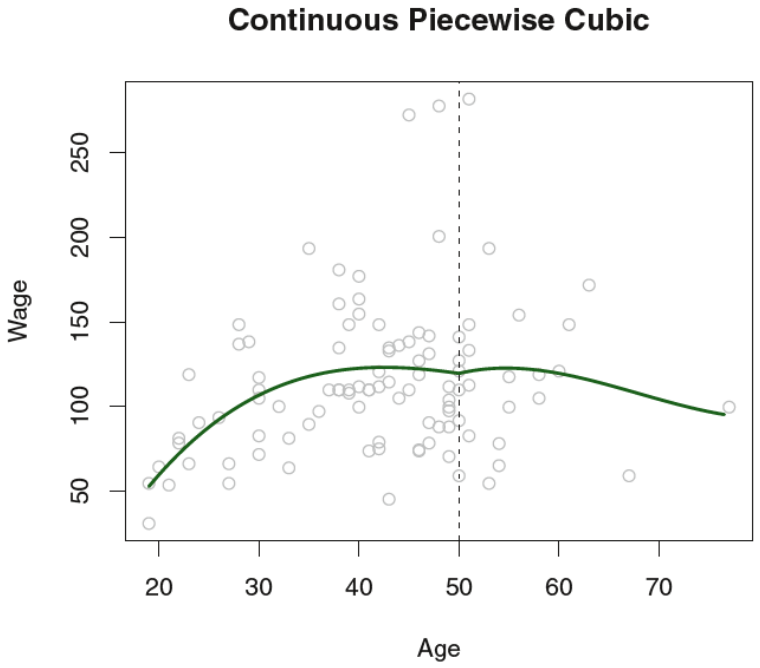
SPLINES

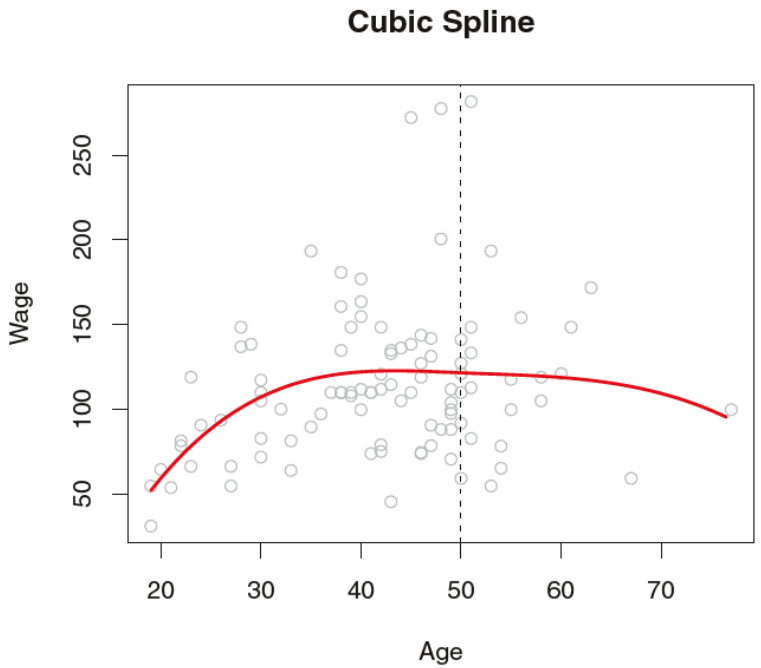
Piecewise polynomial regression

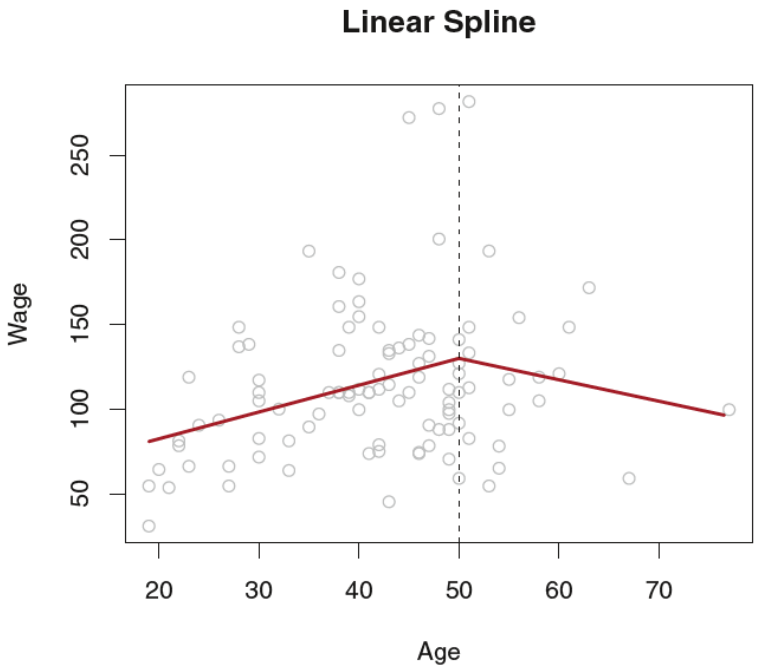
Number of knots

Fitted curve must be continuous…
…first and second derivative of the function need to be continuous
…smooth









REGRESSION TREES

REGRESSION TREES

Decision trees to make predictions for continuous targets

Instead of entropy we want to reduce variance

$$var(t, D) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n - 1}$$

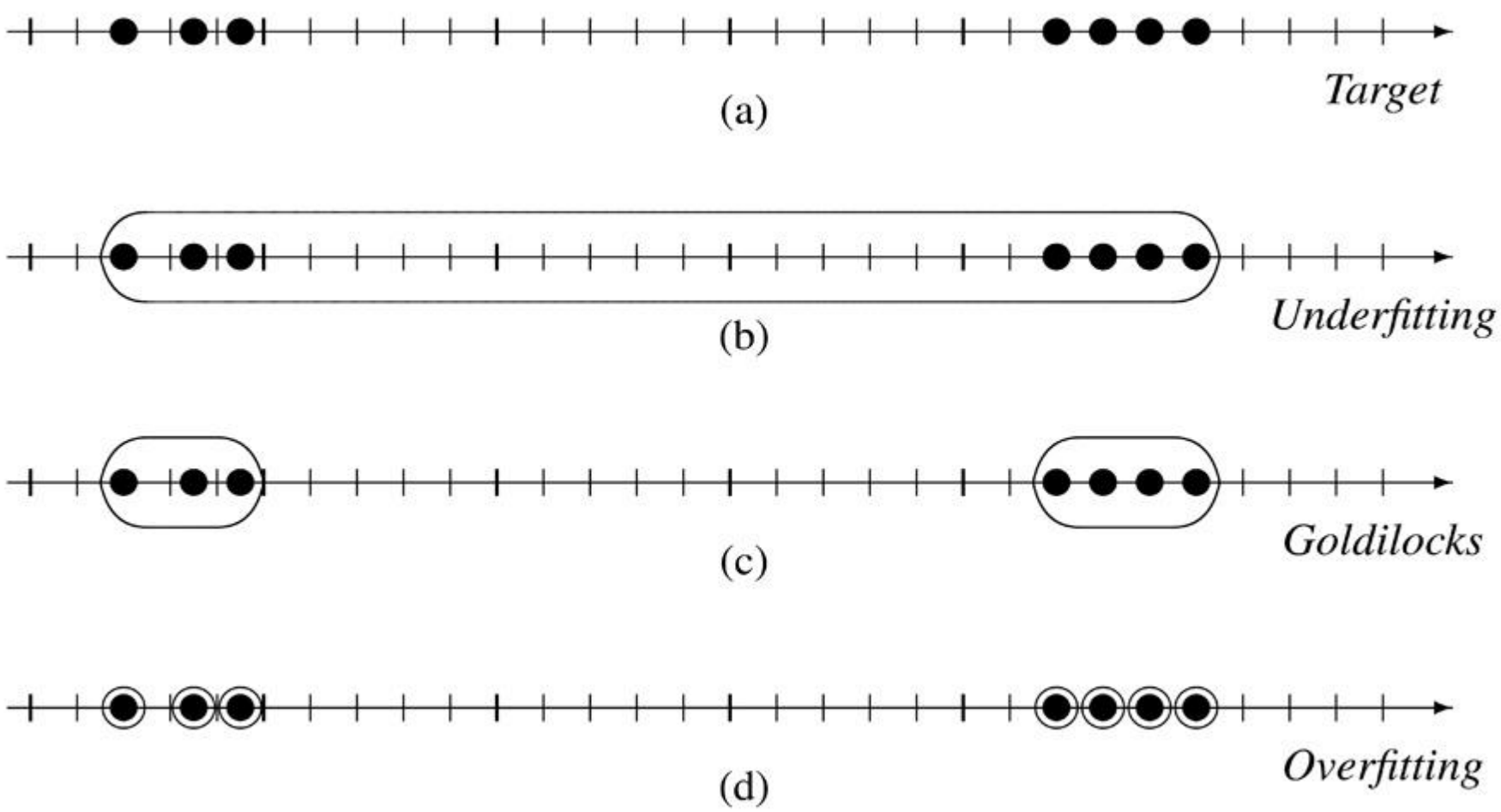
D is a dataset that has reached the node
n is the number of datapoints in D

CHOOSING THE SPLIT

Choosing which feature to split based on summed weighted variance

$$d[best] = \min_{d} \frac{|D_{d=l}|}{|D|} \times var(t, D_{d=l})$$

HOW WE WANT TO SPLIT?



WORKFLOW OF DECISION TREES

Require: set of descriptive features ***d***
Require: set of training instances ***D***
~~**if** all the instances in D have the same target level *C* **then**~~
~~**return** a decision tree consisting of leaf node with label *C*~~
else if ***d*** is empty **then**
return a decision tree consisting of a leaf node with the *y* of the mean of y ∈ ***D***
else if ***D*** is <%5 of all datapoints **then**
return a decision tree consisting of a leaf node with the *y* of the mean of y ∈ ***D*** of the immediate parent node
else ….

WORKFLOW OF DECISION TREES

...

else

d [*best*] <- arg max *InformationGain(d,D)*

make a new node, $Node_{\boldsymbol{d}[best]}$ and label it with ***d*** [*best*]

partition ***D*** using ***d***[*best*]

remove ***d***[*best*] from ***d***

fore each partition $\boldsymbol{D}_i$ of ***D*** do

grow a branch from $Node_{\mathbf{d}[best]}$ to the decision tree with $\boldsymbol{D} = \boldsymbol{D}_i$


ENSEMBLE LEARNING
Ludwig

ENSEMBLE LEARNING

We have seen how to develop a single most accurate prediction model

We can also develop a set of models and make predictions by aggregating the outputs together

COMMITTEE OF EXPERTS

Experts working together are more likely to solve it then individual experts

Guard against group thinking

Ensemble can perform accurately even if individual models do only slightly better then random guessing

ENSEMBLE MODELS

(1) Build multiple different models from the same dataset…
…but inducing each model on a modified version of the dataset

(2) Predictions are made by aggregating the predictions of the different models…
…classification: voting mechanisms
…regression: central tendency (mean, median)

BOOSTING

Each new model is biased to pay more attention to instances that previous models “missed”

We use weighted datasets, each point has weight $w_i \geq 0$
Initially $w_i = 1/n$

Weights are used to create replicated datasets
…the number of times the point is sampled is proportional to weight

BOOSTING

CLASSIFICATION

(1) Induces a model, calculates total error ϵ by summing the weights of missed instances

(2) Increases the weights for misclassified datapoints

$$w[i] \leftarrow w[i] \times \left(\frac{1}{2 \times \epsilon}\right)$$

Or decrease for correct

$$w[i] \leftarrow w[i] \times \left(\frac{1}{2 \times (1 - \epsilon)}\right)$$

(3) Confidence factor of a model

$$\alpha = \frac{1}{2} \times \ln\left(\frac{1 - \epsilon}{\epsilon}\right)$$

GRADIENT BOOSTING

REGRESSION

(1) Train the regression tree
(2) Apply the regression tree just trained to predict
(3) Calculate the residual of this decision tree and save residual errors as the new y
(4) Repeat from (1) until the number of trees we set to train is reached
(5) Make the final prediction

LETS PRACTICE

OIL SAMPLES

1H and 13C NMR spectra were obtained from 126 petroleum samples and values of their total acidy number (TAN), ranging from 0.03 to 4.96 mg KOH· g^{-1}, to distinguish the oil samples from the TAN values.

Anneli Kruve

RASPBERRIES

Considering the short shelf life of raspberry, the processing, storage, and transport are some of the main issues to be addressed.
A comparative experiment was conducted in order to find the suitable process parameters for convective drying that may be considered as the alternatives to freeze-drying, which is a widely used preservation method for raspberry even though it is a costly and energy-consuming method. Twelve convective drying regimens were applied with a combination of three influencing factors: air temperature (60°C, 70°C, and 80°C), air rate (0.5 and 1.5 $m \cdot s^{-1}$), and stage of raspberry (fresh and frozen). .....
..... showed that convective drying of fresh raspberries proved to be more similar to freeze-dried raspberries than convective drying of frozen ones. Fresh samples dried at 60 °C air temperature and 1.5 $m \cdot s^{-1}$ air flow proved to be the most similar to the reference freeze-drying method. This convective regimen gives samples with the lowest color change, shrinkage, and shape deformation and most similar mechanical and chemical properties of the samples.

Anneli Kruve

PROTEIN SECONDARY STRUCTURE

The secondary structure analyses of proteins hold immense importance in the field of protein science because it plays a vital role in its hierarchical classification. It is the most important transitional step in the prediction of the three-dimensional structure of any protein.
Here the accurate secondary structure of FAP174 was investigated through circular dichroism (CD) and Fourier transform infrared spectroscopy (FTIR) combined with …. The far-UV CD spectrum of FAP174 exhibited a positive band at 192 nm and negative bands at 208 and 221 nm. These results were further confirmed by FTIR spectrum of FAP174 that revealed amide **I** and **II** bands at 1654 and 1547 cm^{-1}, respectively. And the …… models led to the quantification of the secondary structure for FAP174 protein that fairly corroborated with the values obtained by quantifying CD spectrum, these being approximately 54% α-helix, 0% β-sheets, approximately 12% β-turns, and approximately 34% other structures.

Anneli Kruve

ACID-BASE PROPERTIES

Quantitative structure-activity relationship models (QSAR models) predict the physical properties or biological effects based on physicochemical properties or molecular descriptors of chemical structures.
Here the p*K*a values of the OASIS database (1912 chemicals) were predicted based on the molecular structure descriptors and by the combination of …….. According to the results, the prediction performance was enhanced by more than 15% (root-mean-square error [RMSE] value) compared with the predictions of the best individual QSAR model.

Anneli Kruve

COSMIC MASS SPECTROMETRY

The instrument COSIMA (COmetary Secondary Ion Mass Analyzer) onboard of the European Space Agency mission Rosetta collected and analyzed dust particles in the neighborhood of comet 67P/Churyumov-Gerasimenko. The chemical composition of the particle surfaces was characterized by time-of-flight secondary ion mass spectrometry. A set of 2213 spectra has been selected, and relative abundances for CH-containing positive ions as well as positive elemental ions define a set of multivariate data with nine variables. Evaluation by complementary chemometric techniques shows different compositions of sample groups collected during two periods of the mission. The first period was August to November 2014 (far from the Sun); the second period was January 2015 to February 2016 (nearer to the Sun).
The applied data evaluation methods consist of ……. The results indicate a high importance of the relative abundances of the secondary ions C^+ and Fe^+ for the group separation and demonstrate an enhanced content of carbon-containing substances in samples collected in the period with smaller distances to the Sun.

Anneli Kruve

BEERS

In this study, 13 properties (alcohol-, real extract-, flavonoid-, anthocyanin, glucose, fructose, maltose, sucrose content, EBC [European Brewery Convention] and L*a*b* color, bitterness) of 21 beers (alcohol-free pale lagers, alcohol-free beer-based mixed drinks, beer-based mixed drinks, international lagers, wheat beers, stouts, fruit beers) were determined. In the first step …… was performed for the whole data and five clusters were determined; then, a bootstrapping was applied to establish a balanced data so as every cluster should contain 100 samples and the total sample size is 500.

Anneli Kruve

GASOLINE

Commercial gasoline must satisfy several product specifications before trading. ……. was applied to create reliable prediction models for 13 physicochemical parameters (e.g, density, vapor pressure, evaporate at 70˚C, evaporate at 100˚C, evaporate at 150˚C, final boiling point, research octane number, motor octane number, aromatic content, olefinic content, benzene content, oxygen content, and methyl tert-butyl ether content) of gasoline produced in Matosinhos' refinery. The input variables for the regression are the 1H NMR spectral intensities of a total of 448 samples, which were recorded using a picoSpin NMR spectrometer operating at 80 MHz. The output variables are the corresponding property values, which were also measured according to ISO standard methods. The optimum complexity of each model was achieved by repeated double cross-validation strategy, consisting of 100 repetitions of two nested cross-validation loops. ……. yielded accurate predictions of 11 of 13 properties within the reproducibility of ISO standards.

Anneli Kruve

ROCK SOLUBILITY

This study uses 888 rock samples collected from the exploration and production (E&P) sector of the oil industry. Based on the Fourier-transform infrared (FT-IR) spectra of these rock samples their solubility predictions have been developed and investigated with ……, ……, ……..
The investigation starts with spectral data pre-processing, baseline correction and feature selection methods creating the feature set for all machine learning (ML) applications. The accuracy of predictions has been evaluated with mean squared error as a performance metric for each investigated method. The comparisons of predicted values to real data of test samples have shown that mineral solubility in acids can be well predicted in the range of the uncertainties of real laboratory measurements, therefore it can be used to improve the response time of these investigations and reduce the risk in industrial applications. In those cases, where the unknown samples have got some out of the range features, the limitations in the accuracy of predictions have become clear. The identified constraint of samples' multitude further emphasizes the need for database building efforts, so that the real potential in big data and machine learning can be realized.

Anneli Kruve

MAGNETIC TAPES

Audio magnetic tapes manufactured using polyester urethane are known to become nonplayable over time due to the degradation of the magnetic layer. Attempting to play degraded tapes to digitize them can cause extensive damage to the tape as well as to the play back device. For this reason, most of the magnetic tapes in cultural heritage institutions are in critical state. The purpose this study is to develop a nondestructive technique to determine degradation status. Our approach is to combine attenuated total reflectance Fourier transform infrared spectroscopy (ATR FT-IR) with chemometric techniques ..... The model 1 built was able to successfully classify playable and nonplayable with 97% to 98% accuracy when similar tape brands/models were in the training and the test set. With different brands/models in the test set, the model 1 performed poorly. However, model 2 showed 95.5% accuracy for similar brand/models and 80.5% accuracy for different tape brands/models.

Anneli Kruve

NIFUROXAZIDE

…… methods were able to determine simultaneously the quinary mixture of nifuroxazide (NIF) and its four carcinogenic impurities. …… models were built covering the range of 10.00 to 50.00 µg mL^{-1} for NIF, 0.05 to 0.45 µg mL^{-1} (for each of impurities A and B), and 0.10 to 0.90 µg mL^{-1} (for impurities C and D). NIF and its four genotoxic impurities were successfully determined in the prepared mixtures and dosage forms. The efficiency of the applied algorithm for resolution and quantitation of the overlapped UV signals were compared.

Anneli Kruve