Affective Computing

Felix Putze
„Design and Evaluation of innovative User Interfaces“
WS 2013/14
Emotion Recognition: Goals

• Motivate relevance of affective computing
• Show applications of affective computing
• Describe steps to build a generic emotion recognizer
• Show that there is more to emotion recognition than just machine learning:
  • ground truth forming
  • data collection
  • feature selection
  • evaluation of systems
• Cowie et al.: Beyond emotion archetypes: Databases for emotion modelling using neural networks, Neural Networks 18(4), 2005
Why Affective Computing? \rightarrow Social Factors

- Media Equation (Nass):
  - Users react to social signals (like emotions) send by the system
  - Users expect systems to social signals they send
  - Machines which do not follow those rules are perceived as unintelligent and impolite
- To optimally adjust to the user, the system must have *empathy*: The ability to sense the feelings of other people
- \rightarrow If we want to design systems with which humans interact naturally, we need to incorporate “emotional intelligence”
Brave, Nass, Hutchinson (2005): Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent

- Self-oriented vs. other-oriented emotion
- Display of other-oriented emotion increased likability and trustworthiness
- Display of self-oriented emotion influenced perceived submissiveness

Example: Affective Blackjack Agent

That’s great! I’m really happy that you won. (other)

I won! That cheers me up! (self)

Oh no, you lost! That’s too bad. (other)

I lost this hand. I feel very sad. (self)
Why Affective Computing? → Cognition

- Emotion and affect dramatically influence human cognition
  - Positive influence: motivation, attentiveness, activation, creativity, ...
  - Negative influence: tunnel vision, fatalism, unjustified pessimism/euphoria, ...

- Psychology and Neuroscience indicate that emotion is a crucial factor for decision making
  - Influence learning, adaptation
  - Emotional responses often much faster than rational ones
  - Emotion-impaired people have difficulties in making rational decisions

- Emotions are not only relevant for the “feel good factor” but also for efficiency and effectivity of a system
  - See extended example later (AutoTutor)
Affective Systems

- Recognizes the user’s emotional state using sensors or context
- Adapt to the user’s emotional state, for example:
  - Mimic the user’s state
  - Calm down or motivate the user
- No recognition of true emotional state but recognition of emotions as communicated (as humans do)
  - We cannot look into the user’s mind, but we can use all external queues available to resolve ambiguities
- Not limited to emotion but can target other affective or cognitive states
  - Mental workload
  - Non-emotional social signals (dominance, cooperativeness, ...)
  - Personality
Emotion: Definition

- No single accepted definition
- Most definitions agree on the following:
  - Subjective experience of one’s state of mind
  - Universal among all humans
  - Coupled with physiological changes of the body
  - Short duration (a few seconds max)
  - Triggered by a certain stimulus
- Differentiate emotion from related concepts
  - Attitude (evaluation of a concept in general):
    - “I find it uncomfortable to go to the dentist.”
  - Mood (longer lasting, general feeling):
    - “I am nervous because I have to go to the dentist tomorrow.”
  - Emotion (short-lived, stimulus-triggered feeling):
    - “I just heard someone scream in the doctor’s room. I am frightened!”
Emotion Models

• Need model to represent fuzzy concept „emotion“ computationally
• Used to formalize which types of emotion the system can handle and distinguish
• Crucial step for designing a system:
  • What is left out here cannot be represented by the system
  • System needs to recognize and react to all states defined by the model
• Model for a specific HCI application, model may contain specific affective concepts
  • engagement, vigilance, stress, attention, ...
• More on affective models → See lecture on Cognitive Modeling in summer term!
Emotion Models: Categorial Emotions

- Small set of basic emotions
  - e.g. anger, fear, disgust, surprise, joy, sadness, etc. (according to Ekman)

- Characteristics:
  - Distinctive universal signals (e.g. facial expressions)
  - Cross-culturally displayed and recognized
  - Innate, not learned
  - Well discriminated by
    - Physiological responses
    - Associated reaction patterns
  - Designed by experts

- Problem:
  - Inflexible model
  - No representation for blended or alleviated emotions
Emotion Models: Dimensional Model

- Describe a small number of dimensions
- Emotion is point in this multi-dimensional space
  - Russell (1980): Bipolar circumplex model (dims: Valence and Arousal)
    - Valence (quality): unpleasant to pleasant
      - Positive vs. negative affective states
    - Arousal (quantity): calm to excited
      - Mental and/or physical activity level
    - Dominance (control): weak to strong
      - Control or lack of control over others or situations
- Mostly generated from data corpora (data-driven approach)
  - Emotional adjectives, Facial expressions, Emotional experiences, ...
  - Statistical methods to find latent dimensions in data corpora
Mapping between Emotion Models

arousal

surprise

elated

happy

pleasure

displeasure

afraid

stressed

frustrated

sad

neutral

depression

bored

content

calm

relaxed

sleepy

sleep

12/72
Emotion Models: Appraisal Models

- Appraisal models: Characterize emotions as evaluations of events with regard to their goals and attitudes

  1. Relevance Detection ("Is the event new and of importance?")
  2. Implication Assessment ("Is event helping or obstructing my goals?")
  3. Coping Potential Determination ("Can I influence outcome of event?")
  4. Normative Significance Evaluation ("Is event in accordance with the attitude and morality of myself and others?")

- Example: Student succeeds in a hard exam
  → Unexpected, helping the student’s goals, influenced by student’s learning effort, creates good reputation with others
  → Feeling of pride

- In other circumstances (e.g. student with big ego, i.e. success not unexpected), resulting emotion is different
Transmitting Emotions

- Emotions are transmitted to other people...
- ...with communicative function
  - To indirectly inform about the cause of the emotion (e.g. fear to indicate the presence of something dangerous)
  - To reveal information about the inner state (e.g. to create appropriate responses)
  - To fit in the norm (e.g. one is expected to express sadness on a funeral)
- ...as by-product of preparing the body for typical responses
  - flight in the case of fear → prepare the body for running
Modalities conveying Emotions

- **Speech/Language**
  - Word choice
  - Voice tone
  - ...

- **Motion**
  - Facial expressions
  - Gestures
  - Posture
  - ...

- **Physiology**
  - Skin temperature
  - Respiration
  - Muscle tension
  - ...

I feel good!
Voice Analysis

+ captures emotion communicated by speech and language
+ contains low-level information (tone of voice)
+ contains semantic information (if speech recognition is available)
+ requires only a microphone as sensor

- only available when user speaks
- might violate privacy of conversations
- easily controllable by user
Video Analysis

- captures emotion communicated by motion
- captures large scale of expressions
  - small facial movements
  - hand and arm gestures
  - Whole body pose, walking style
- Always available (if recording already in place)

- requires camera to be pointed at the user (difficult in outdoor situations)
- dependent on lightening condition
- easily controllable by user
- reveals identity
Physiological Signal Analysis

- + hard to control by the user
- + nearly always available
- + mobile

- strong signal variance across conditions
  - Sensor positioning
  - Temperature
  - ...

- requires attachment of sensors (but those are small → wearable computing)
Examples of Physiological Signals

Signal for “Anger”

Signal for “Grief”
Context

• Context plays an important role in emotion assessment
• Context information can be used as information source
  • To evaluate a person’s appraisal of an event
  • Reaction to a perceived emotion without knowing the context might cause trouble (inappropriate reaction to perceived “anger” annoys user even more)
• Context may limit or influence recorded signals
  • No gestures when hands are occupied
  • No video recording at night
  • Whispered speech in the library
  • Difficulties with physiological sensors in the heat
  • All modalities: Other activity influences the recorded signals
    – Physical activity
    – Non-emotional aspects of communication (cultural, social, environment-specific, ...)

DEiB: Affective Computing
Ethical Considerations

• Do we actually want to discover emotions user’s intentionally want to hide?
  • People may control their information despite the benefit of an emotion-adaptive system
  • Is reacting to emotions people are hiding incapacitation of the users, even if done with best intentions?

• Answer depends on the system we are building
  • Probably unjustified for purely usability-oriented systems
  • Maybe justified for a system to support user’s in stressful, safety critical tasks

• Answer depends on the reaction of the system
  • Passive feedback to the user is less problematic than...
  • Autonomous reaction or reinterpretation of user’s requests
Application: AutoTutor

- How to integrate an automatic emotion recognizer into an application?
- AutoTutor: Automatic learning assistant
  - Learning sessions in natural language
  - Appears as an animated agent
  - Supports long sessions (>100 turns)
- Integration of affect recognition
  - Based on conversational features, posture, facial expression and physiological signals
  - Models boredom, confusion and frustration
AutoTutor: Attribution Theory

- Attribution theory: Describes how people explain their success or failure based on three features
  - Internal vs. external
  - Stable vs. unstable
  - Controllable vs. uncontrollable

- Attribution influences how learning will progress in the future

- People tend to view themselves positively:
  - Success is attributed to stable, internal, controllable factors
  - Failure is attributed to external, uncontrollable factors

- Learner should attribute failure to internal, stable, controllable factors by setting manageable learning goals

- Can be reached by showing empathy for the learner
  - Increases influence of tutor on learner
  - Helps to check whether goals are still manageable (or too easy)
AutoTutor: Cognitive Disequilibrium

- Cognitive disequilibrium: Discrepancy between belief and observations, e.g. contradictions or surprises
- Entering a state of cognitive disequilibrium usually activates cognitive deliberation
  - → Learner is interested and motivated
- Staying in cognitive disequilibrium for too long creates a feeling of confusion and frustration
  - → Learner might give up
- Emotion recognition helps the tutor to determine whether learner is in state of “good” cognitive disequilibrium
AutoTutor: Affective Responses

- AutoTutor contains a set of rules to react to affective cues
- Mapping: cognitive and affective state → tutor actions
- Parameters describing learner’s state:
  - Current emotion detected
  - Confidence level of the classification result
  - Previous emotion detected
  - Student’s ability (measured dynamically from his answers)
  - Quality of the student’s last response
- Example:
  - Overall good student
  - Last answer was not satisfactory
  - Current emotion classified as boredom (high confidence)
  - Previous emotion classified as frustration
  - “Maybe this topic is getting old. Let’s finish to try something new.”
AutoTutor: Affective Responses

- Short feedback: from positive to negative depending on answer quality
  - Short statement (e.g. “well done” or “you are on the wrong track”)
  - Facial expressions (e.g. showing approval or skepticism)
  - Not emotion-adaptive

- Feedback is followed by an emotional response if the system senses the need to motivate the student. Employed communication modalities:
  - Content of the feedback statement
  - Facial expressions
  - Emotionally modulated speech

<table>
<thead>
<tr>
<th>Affective State</th>
<th>Pitch Range</th>
<th>Pitch Level</th>
<th>Speech Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surprise/Delight</td>
<td>Wide</td>
<td>Very High</td>
<td>Fast</td>
</tr>
<tr>
<td>Empathy</td>
<td>Narrow</td>
<td>Low</td>
<td>Slow</td>
</tr>
<tr>
<td>Skeptical</td>
<td>Narrow</td>
<td>High</td>
<td>Slow</td>
</tr>
<tr>
<td>Disappointment</td>
<td>Narrow</td>
<td>Low</td>
<td>Slow</td>
</tr>
</tbody>
</table>
AutoTutor: Response Strategies

- **Supportive strategy: empathetic and motivational responses**
  - Attribute source of the learner’s emotion to the material or itself instead to learner
  - Boredom → “Let’s keep going so we can get to the interesting stuff.”
  - Confusion → “I know I do not always convey things clearly.”
  - Frustration → “This material is difficult but I know you can do it.”

- **Shakeup strategy: attribute emotions to the learner**
  - In accordance with results of attribution theory
  - More conversational, colloquial style
  - Confusion → “This material got you confused but I think you can do it.”
  - Boredom → “Geez, this stuff sucks. I’d be bored too.”

- Other experiments suggest that preference of strategy depends on the user’s personality (might also be worth determining in an affective system)
According to Feigh (2012)

- Not only applicable to affective computing but to all kinds of adaptive interfaces

- e.g. pausing when arousal is too high or too low
- e.g. changing appearance or voice of avatar
- e.g. avoiding distressing content when negative affect detected
Challenges of Adaptive Systems (1)

- Jameson (2009) lists a number of challenges and side effects of adaptive computation that need to be addressed
  - Need to Switch Applications or Devices
    - Needs more sensors, computing capabilities, network connections, ...
    - User needs to invest time and money
  - Need to teach the system
    - User needs to provide sensor data or other information for full use of system (e.g. rollout of recognizer, ...)
  - Narrowing of experience
    - Adaptive system may hide certain “states of the world” from the user (e.g. if content is
    - Removes opportunity for experience and learning
Challenges of Adaptive Systems (2)

- **Unsatisfactory Aesthetics or Timing**
  - Automatically generated content, timings or appearance may be inferior to thoughtfully hand-crafted content and behavior

- **Need for Learning by the User**
  - Adaptive system might show unexpected behavior compared to conventional systems
  - Capabilities and limitations unknown
  - User might need to learn new behavior to benefit from adaptive system (e.g. give up some control and let the system take some decisions)

- **Inadequate Control over Interaction Style**
  - Adaptive behavior may collide with user’s preferences (adaption replaces adaptivity)
  - Requires feedback mechanisms to report undesired behavior
Challenges of Adaptive Systems (3)

- Threats to Privacy
  - Data is collected and interpreted
  - Private information may be distributed to third parties
  - Preferences or attributes might be revealed to the public (or user)

- Inadequate Predictability and Comprehensibility
  - System behavior becomes indeterministic and dependent on hidden variables $\Rightarrow$ increase in complexity
  - User cannot plan or optimize behavior

- Imperfect System Performance
  - Statistical models will sometimes provide misinterpretation of data
  - A few „stupid“ decisions of the system might outweigh a majority of useful decisions
Data Collection

- Want to build statistical models of how physiological signals correspond to affective states
- Need representative data to train and evaluate the models!
- Requirements to a good data set:
  - High data quality (complete data sets, low noise, ...)
  - Actually contains the desired effects (→ experiment design)
  - Realistic (i.e., can generalize to the real world)
  - For person-independent systems: Cover a representative group of potential users
Methods of Generating Emotional Data

• Acting
  • Generates data with many of full-blown emotions
  • Easy to discriminate, helpful to test new methods
  • Produced signals may be completely different from real ones (e.g. controlled smile uses other muscles than a genuine one; take the test here: [http://www.bbc.co.uk/science/humanbody/mind/surveys/smiles/index.shtml](http://www.bbc.co.uk/science/humanbody/mind/surveys/smiles/index.shtml))

• Emotion Elicitation
  • Create situations in which people experience the desired emotion
  • Still controlled, but emotions usually genuine
  • People may react differently (or not at all) to the stimuli

• Real World Emotions
  • Most representative
  • Privacy issues
  • Real emotions are rare and usually ambiguous (see: Douglas-Cowie et al., 2000: A new Emotion Database: Considerations, Sources and Score)
**Example: Acted Emotional Speech**

- **Berlin EmoDB (Burkhardt et al., 2005)**
  - 6 basic emotions, neutral semantic content
  - Professional German actors (5 female and 5 male)
  - Clearly distinguishable using phonetic and acoustic markers
  - Super unrealistic

- happy
- angry
- frightened
- bored
- sad

- sad

- ### Berlin EmoDB (Burkhardt et al., 2005)

- 6 basic emotions, neutral semantic content
- Professional German actors (5 female and 5 male)
- Clearly distinguishable using phonetic and acoustic markers
- Super unrealistic

- happy
- angry
- frightened
- bored
- sad
International Affective Picture Set

• There exists a small number of standardized methods for emotion elicitation
• The International Affective Picture Set (IAPS) contains a large number of photos which are labeled by hundreds of people of different nationalities for emotional content
• Different labels for men and women (men rate sexual content with high valence, women sports and babies)

For illustration only. IAPS pictures may not be revealed to the public to not compromise its validity. The low valence pictures also get much more explicit than this example.
Five Factors in Eliciting Emotion

- There are many methods for emotion elicitation available
  - Presenting films, music, or pictures
  - Have subject make faces (for elicitation, not acting)
  - Creating emotional interaction situations
  - ...

- Discriminate them on five dimensions:
  - Subject-elicited vs. event-elicited (e.g. self-imagined emotional situations vs. stories read by experimenter)
  - Lab setting vs. real-world
  - Expression vs. feeling (emphasis on external expression or on internal feeling?)
  - Open-recording vs. hidden-recording (do subjects know their data is recorded?)
  - Emotion-purpose vs. other-purpose (do subjects know experiment deals with emotion?)
Natural Emotions

- AIBO corpus (Batliner, Steidl et al., 2004)
  - Children interacting naturally with the AIBO robot
  - Participants controlled robot dog through parcour
  - No instructions on the speech commands they could use
  - Children less disguise their emotions compared to adults
  - Wizard-of-Oz controlled behavior: Obedient or disobedient
  - Very variable reactions to similar situations
Generating Ground Truth

- For training statistical models, we need to assign ground truth labels to the data, according to the employed model.

- A-Priori assignment
  - Setting is designed to create the emotion „anger“ → label it as such.
  - What if we are wrong about the effect of our treatment?

- Rating by Judges
  - Have (naïve or expert) judges watch/listen to the data and rate it.
  - Relies on humans as emotion experts.
  - Judges cannot peek insight a subject’s mind.

- Rating by Subjects
  - Have subjects label the data themselves.
  - Yields their true subjective emotions.
  - Approach 1: During the experiment, e.g. think-aloud (intrusive).
  - Approach 2: Label afterwards (memory might betray subjects).
Labeling Protocols

- A protocol defines how to assign labels to emotion
- Design decisions for a labeling protocol
  - What is the underlying emotion model?
  - Which emotions/dimensions are annotated?
  - Annotate strength of emotion?
  - Annotate certainty of labeler?
  - Allow blended emotions (multiple emotions at the same time)?
  - Provide fallback classes like “non-neutral”, “other” or free input?
- Also related to the employed annotation language
  - Example: Emotion Markup Language (http://www.w3.org/TR/emotionml): XML-based generic markup to support different models (categorical, dimensional, appraisal-based)
Examples for Labeling Tools

Feeltrace tool for time-continuous annotation of videos according to a two-dimensional emotion model (valence, arousal)

Self-assessment Manikin (SAM) for rating emotions in a three-dimensional emotion model (valence arousal, dominance) on a questionnaire
Selection of Classes

- Most systems use a finite set of classes for labeling.
- Classes should be balanced, i.e. be of approximately same size in the data set.
- Imbalance might require collapsing classes (e.g. “disgust” + “anger” + “sadness” = “negative”).
- Classes should be defined by the “emotional space” of the application.
  - How many HCI applications deal with the emotion “disgust”?
  - Applications in a learning context might deal with the emotions “boredom” and “confusion” (although no basic emotions).
  - Applications for children might deal with the emotion “motherese” *)
- If you are using a continuous dimensional model, either:
  - Discretize data (e.g. form “high valence” and “low valence” classes)
  - Perform regression instead of classification (i.e. keep continuous values)

*) see Stefan Steidl: Automatic Classification of Emotion-Related User States in Spontaneous Children’s Speech, 2009
Inter-Labeler Agreement

- People make mistakes in assigning labels and data is ambiguous → increase reliability by having multiple labelers.
- How to handle disagreement between labelers?
- Combine multiple labels by
  - majority voting (discrete classes)
  - averaging (continuous scale)
- Strong disagreement can also be an indication of
  - Problems with the data segment (does not contain any emotion in scope of application)
  - Problems with the emotion model (might not cover all relevant emotions)
  - Instructions of labelers (might not know how to rate emotions)
Measuring Agreement

- Can we measure how well a group (here: 2) of labelers agrees?
- First try: measure agreement as overlap of labels in %
- Problem: Does not respect imbalanced classes or number of classes → numbers are not comparable across experiments
- Example: An agreement of 70% is high for four classes of equal size but low for two classes of which one contains 70% of all samples

- One solution is Cohen’s Kappa: \[ \kappa = \frac{p_0 - p_c}{1 - p_c} \]
- Here, \( p_0 \) is the actual agreement between both labelers, \( p_c \) is their agreement by chance
- Values above 0.6-0.7 are usually considered “good”
Existing Databases

- Several reasons for employing published emotion databases:
  - Data collection and labeling is expensive and time-consuming
  - Allows comparison of different algorithmic approaches
  - Allows to test generalizability of an approach on rich data sets
- Drawback: May not fit perfectly to envisioned application
- Collection of databases: http://emotion-research.net/databases (also contains other tools and resources for emotion research)
- Example: “Vera am Mittag” – Corpus
  - Recorded from a german talk show
  - Movies, stills and audio of several dozen subjects
  - (Mostly) naturalistic emotions
- There are quite many emotional databases available containing speech and video. Physiological databases are harder to extract from TV or meeting recordings
Feature Definition

- A feature is an observable variable explaining the dependent class attribute
  - Example: Pulse Signal → Heart Rate, Heart Rate Variability
- Features are the input to the classifier during training and testing phase
  - In many cases, selection of good features has much higher impact on result than selection of classifier
- Features can be synthetic or knowledge-based
  - Synthetic: Statistical moments, Frequency characteristics, temporal patterns, ...
  - Knowledge-based: Signal- and/or application specific signal attributes
Inference and Validity

- Fairclough (2009) defines a number of challenges for physiological computing, among others:
  - Psychophysiological inference (relationship between physiological measurement and psychological construct)
  - Psychophysiological validity

- The relation between construct (e.g. emotion) and measurement (e.g. heart rate)
  - should be one-to-one... (i.e. “isomorphic”)
  - ...but is often many-to-one (e.g. there are many constructs which have the same influence on the measurement) or many-to-many
  - ...should be context-independent (e.g. transferrable from laboratory to the field)

- As a high level of diagnosticity may be difficult to achieve, the designer has to decide accordingly (e.g. is it acceptable if stress and anger are confused?)
Inference and Validity

• A good measurement needs to have a high sensitivity
  • Can we only detect presence/absence of a certain state (e.g. anger vs. no anger) or can we differentiate finer levels?

• A good measurement needs to have a high generalizability
  • Low influence of intra-/inter-subject variability

• A black-box machine learning algorithm learns to discriminate two experimental conditions, not necessarily two psychological states
  • Necessary to validate the system with a number of concurrent criteria
    – exposure to media
    – experimental tasks
    – subjective measures
    – observable behavior
Example for Thread to Validity

- **Task:** Discriminate low and high arousal
  - Perform emotion elicitation with film clips
  - Low arousal: calm, quiet nature scenes
  - High arousal: horror and action films

- Pulse, Respiration frequency and skin conductance show good discriminative power between low and high arousal condition.

- **BUT:** There are many other differences between the sets of film clips:
  - Nature scenes clear and bright, horror films dark and murky
  - Differences in the number of scene changes and cuts
  - Differences in sound characteristics
  - ...

- **Rule of thumb:** The more realistic the scenario, the higher the number of potential differences
Other Attributes of a good Feature Set

- Features can be extracted reliably
  - Invariance properties (e.g. invariant to variation in time, space, ...)
  - May require normalization, transformation, ...
- Features have a high signal-to-noise-rate
  - More robust models with less training data
- Features are not redundant
  - Redundant features may bias statistical models
  - One-to-many relation may indicate low diagnosticity
- Features are commensurable
  - More robust model estimation
- Features are fast to calculate
  - For real-time applications
Curse of Dimensionality

- With the number of features, the complexity of statistical models increases
  - Large number of parameters may only be estimated from sparse data
  - Complex model is capable of overfitting

- Rule of thumb: #training samples $> 10 \times$ feature dimensionality
Feature Selection

• Often, one has a large set of potential features to start with
  • Many of those are interrelated
  • Some may not contain the anticipated relation to the ground truth

• We have to select a subset of the most relevant features
  • Manual feature selection methods
  • Wrapper based methods
  • Filter based methods

• Often, combinations of all those methods are employed to improve the quality of the final feature set
Visual Inspection

- Scatter Plot: Shows distribution of samples as a function of two features
- Can be employed to check whether classes are separable using those features
- Can be employed to check whether two features are interrelated (e.g. one grows with the other)
- Helps to identify outliers in the data

- Scatter Matrix: All combinations of pair wise scatter plots to display a set of features
Filter Based Feature Selection

- Filter based methods are automatic but classifier-independent
  - May or may not use class information
- Two main approaches exist:
  - Calculate statistical properties of a feature to rank the features:
    - correlation coefficient between a feature and the ground truth as a measure of predictive power (positive criterion)
    - correlation between feature candidates as a measure of redundancy (negative criterion)
  - Calculate a projection from the full feature space to a new space of lower dimensionality (e.g. Principal Component Analysis, ...)
    - Example: Linear transformation of 3D coordinates of planar gestures to a 2D subspace
Wrapper Based Feature Selection

- Wrapper based methods incorporate the classifier in the selection process
- Select features by how good they improve the classification
- Possibly better suited for the classifier than filter based selection but more prone to overfitting to the training data
- Typical example: Forward Feature Selection:
  - Initialize tentative feature set $F = \{\}$
  - For all features $f$ not already in $F$:
    - Train a classifier using $F+\{f\}$ as features
    - Test the classifier
  - Set $F = F+\{f\}$ for the $f$ which gave the best classification performance
  - Repeat until a fixed number of iterations is reached or until the classification result does not improve anymore
Evaluation

- After training a classifier, we need to estimate its performance
  - How well does it discriminate classes?
  - How reliably can we estimate performance?
- We are interested in realistic estimates of performance on real-world data
  - Define evaluation metrics
  - Design adequate data sets for evaluation
  - Apply appropriate evaluation schemes
  - Present results appropriately
Quality Measures

- **Accuracy**: Expected percentage of correctly classified samples
- Also called recognition rate
- Typically, the true accuracy of a classifier is unknown and can only be estimated on a given finite data set:

\[ \hat{a} = \frac{\text{#correctly classified samples}}{\text{#all samples}} \]

- Quality of this estimate depends on several factors:
  - How large is the standard error (\( \rightarrow \) depends on variance of accuracy and amount of test data)?
  - How representative is the test data for the whole population?
Precision and Recall

- In many cases, we want to detect a rare positive event in a set with mostly negative samples
  - Detect a specific person in a train station
  - Detect earthquakes in years of seismographic data
- Tuning towards high accuracy would bias the classifier towards labeling every example as negative
- Define other quality metrics: Precision and recall
- Recall:
  \[ \text{recall} = \frac{\text{#correctly classified positive samples}}{\text{#all positive samples}} \]
- Precision:
  \[ \text{precision} = \frac{\text{#correctly classified positive samples}}{\text{#all samples classified as positive}} \]
- F-Score:
  (harmonic mean of precision and recall)
  \[ F - \text{score} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \]
A ROC curve (Receiver Operating Characteristic) curve displays precision and recall for different configurations of the system (e.g. different thresholds).

- Allows to find a tradeoff between both values:
  - High precision for security-critical system
  - High recall for a more usable system

Area under Curve (AUC) is a measure of general classifier performance.
A confusion matrix lists counts of all classes and how they were classified by the system.

- Allows to identify typical confusion pairs (e.g., classes which are very similar) → helps to identify shortcomings of the classifier.

- Identify “sink” classes (which are classified to frequently; happens for example if a classes are imbalanced).

```
row = true label, column = classification results
```

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td>B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td>C</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>E</td>
</tr>
</tbody>
</table>

59/72
Training Set and Test Set

- Evaluating on the same data as used for training will lead to too optimistic results
- Divide available data into disjoint training set and test set, use the latter only for a final test run to estimate accuracy
- A test set may only be touched once to avoid adjusting the classifier towards the data
- Being disjoint may not be enough to guarantee independence of both data sets
  - Multiple sessions by one subject should all be either in training or in test set (otherwise, knowledge on test subjects leaks into training)
- Ensure that test and training set contain (roughly) the same distribution of classes
  - → stratification
Evaluation Set

- A data set which is used in an evaluative context but is used more than once is called evaluation set, development set or tuning set
  - Wrapper based feature selection
  - Tuning of meta-parameters (e.g. learning rates, number of states, ...)
  - Creative experimentation of developer
- Also belongs to the development process, not to the final evaluation
- Accuracies and other scores calculated on this set are too optimistic and should not be reported as final results
- Distribution of data depends on
  - Training set must be large enough to get nearly optimal performance
  - Test set must contain enough data points to get a reliable estimate of the true accuracy
Cross-Validation

- Recording emotional data expensive → small corpora
- Not enough data to create many independent test sets
- Compromise → k-fold Cross-Validation:
  - For a data set $D$, create a number of folds $F = \{ D_1, D_2, \ldots, D_k \}$ with each $D_i$ subset of $D$
  - Iterate over all folds $D_i$ in $F$: Use $F \setminus D_i$ as training set to train the classifier, use $D_i$ as evaluation set to calculate accuracy score $a_i$
  - Average all $a_i$ to report a final accuracy
- Important special case: Leave-one-out cross validation
  - Each fold consists of a single data point for evaluation
- Stratified cross validation: Balance all folds such that the distribution of classes is equal to the distribution in the whole data set
- Note that variance can be large, esp. if $k$ large compared to $|D|$
Cross Validation: Example

Arbitrary 4-partition of all available data *)

Train on partitions 2-4, test on partition 1
→ Estimated accuracy of 81%

Train on partitions 1,3,4, test on partition 2
→ Estimated accuracy of 78%

Train on partitions 1,2,4, test on partition 3
→ Estimated accuracy of 75%

Train on partitions 1-3, test on partition 4
→ Estimated accuracy of 82%

→ Estimated averaged accuracy: 79%

*) Alternatively, we can do random sampling instead of partitioning
Comparison of Systems

- An accuracy score depends on many factors:
  - Quality of classifier
  - Number of classes
  - Quality/composition of test set
  - Intrinsic difficulty of task

- To determine whether a given score is “good”, we need to compare it to another system on the same data set.

- For an established classification task (e.g. speech recognition), this often is an older productive system:
  - Ideally, a published, well-documented system trained on a standardized, published data set
  - More often, our own previous system (be aware of bias towards the new system!)

- For other tasks, we need to find some kind of baseline comparison.
Box Plots

- Box Plots are a good visual tool to summarize performance of multiple runs (e.g. on different test subjects, crossvalidation folds, ...) and also for many purposes...

![Box Plot Diagram]

- median
- mean
- 25% quantile
- 75% quantile
- standard deviation, standard error, ...
- outlier
How to report results

- Typically, a system...
  - has a number of free parameters which can be adjusted
  - contains a number of tuning steps performed

- When evaluating the system...
  - we should be looking at the best combination of parameters
    (not advisable to include mediocre systems in the evaluation)
  - we want to know the sources of largest improvement

- Exhaustive search through parameter space is too expensive

- Alternative:
  - Start with a baseline system
  - Iteratively add one improvement/tune one parameter
  - Does not reveal all redundancy or interaction effects between development steps

<table>
<thead>
<tr>
<th>System</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>71%</td>
</tr>
<tr>
<td>+ Data Normalization</td>
<td>+3%</td>
</tr>
<tr>
<td>+ Feature Selection</td>
<td>+8%</td>
</tr>
<tr>
<td>+ Parameter Optimization</td>
<td>+5%</td>
</tr>
</tbody>
</table>
Relative and absolute Numbers

- If we report an improvement in accuracy from system 1 to system 2, we can report absolute or relative differences
  - Absolute: $\text{acc}_2 - \text{acc}_1$
  - Relative: $(\text{acc}_2 - \text{acc}_1)/\text{acc}_1$

- Note that accuracy scores are always < 100% → relative improvement is always higher than absolute improvement

- Sometimes recognition error is reported instead of accuracy
  - $\text{error} = 1 - \text{accuracy}$
  - For accuracies > 50%, relative decrease of recognition error is higher than relative improvement of the corresponding accuracies
Baseline Systems

- A baseline system is a “simple” system for comparison
  - Employed if no real system for comparison is available
- The simplest baseline system is randomly guessing and produces an accuracy of chance level
  - For c classes, this system has a recognition accuracy of 1/c
- For a very unbalanced class distribution, we can define a better baseline which always guesses the majority class
  - For majority class which covers p% of all samples, the recognition accuracy of the baseline system is p
- For tasks with human-defined ground truth (e.g. emotion recognition), we can use human performance as comparison
  - Not a baseline, but a gold standard (cannot expect to beat humans)
  - Use kappa statistic to estimate how good humans agree on the task
Significance Testing

• If we compare two systems with estimated accuracy scores, we need to check if the difference between both is significant
  • → We can perform statistical tests (e.g. a t-Test)

• For testing, we need a sample of accuracy scores
  • Samples must be independent
  • Data must be interval scaled
  • Sample must be of sufficient size

• How to define the testing sample?
  • Accuracy on a single data sample is not interval scaled (0 or 1) and we cannot assume normal distribution of accuracy
  • Averaging over all data sets does yield only a single accuracy
  • → Provide partition into disjoint testing sets of >30 samples
Confidence Intervals

- Accuracy can only be approximated on the available test data
- Can we estimate a confidence interval for the true accuracy?
- We estimate the probability that a classifier with (unknown) accuracy \(a\) recognizes \(k\) of \(n\) samples correctly (binomially distributed):
  \[
p = \binom{n}{k} a^k (1 - a)^{n-k}
\]
- \(k\) and \(n\) are known for the test set
- \(\rightarrow\) Calculate confidence interval \([a_1, a_2]\) for true accuracy which contains 95% of the probability mass
Bootstrapping

- The former method for confidence intervals does not yield very tight bounds
- We can expect better results using Bootstrapping:
  - Random sampling with replacement from the test set
  - Perform evaluation on each sampled subset
- Calculate the 2.5%-quantile and the 97.5%-quantile of the resulting accuracy values to receive a 95% confidence interval
- Example: After sampling and evaluating, we end up with the following sorted list of accuracy scores:
  [56%, 61%, 67%, ..., 76%, 80%, 84%]

\[\text{95\% confidence interval for true accuracy: [67\%, 76\%]}\]
- Bootstrapping is a very powerful statistical tool with a wide range of applications (and pitfalls to be aware of)
Interpretation of Results

- What to do after all numbers are calculated?
- Summarize strengths and weaknesses of classifier
  - Compare parts of the test set which lead to good/weak performance
  - Inspect confusion matrix
- Evaluate system in context of envisioned application
  - Is recognition rate good enough for typical use cases?
  - Is the tradeoff between precision and recall adequate?
  - Is the recognition fast enough? (else, trade some performance for additional speed)
- Not all errors are equally severe! For example, in ASR:
  - Ground truth: “Please open the documents from the last meeting”
  - Hypo 1: “Briefly open document for last meeting”
  - Hypo 2: “Please delete the documents from the last meeting”
  - Which hypothesis has more errors? Does this number correspond to the severity of error?