Lecture 21: Classification (cont)
Menu for Today (March 21, 2019)

Topics:

- Scene Classification
- Bag of Words Representation

Readings:

- Today’s Lecture: Forsyth & Ponce (2nd ed.) 16.1.3, 16.1.4, 16.1.9
- Next Lecture: Forsyth & Ponce (2nd ed.) 17.1–17.2

Reminders:

- Assignment 5: Scene Recognition with Bag of Words is out
- Midterm solutions are published on Piazza
Today’s “fun” Example: AlphaGo

Google DeepMind’s AlphaGo
Today’s “fun” Example: AlphaGo

Starting out - 10 minutes of training

The algorithm tries to hit the ball back, but it is yet too clumsy to manage.
Lecture 20: Re-cap

- Collect a database of images with labels
- Use ML to train an image classifier
- Evaluate the classifier on test images

Label → Feature vector computed from the image

Example training set

<table>
<thead>
<tr>
<th>Label</th>
<th>Example images</th>
</tr>
</thead>
<tbody>
<tr>
<td>cat</td>
<td></td>
</tr>
<tr>
<td>dog</td>
<td></td>
</tr>
<tr>
<td>mug</td>
<td></td>
</tr>
<tr>
<td>hat</td>
<td></td>
</tr>
</tbody>
</table>
Let $c$ be the class label and let $x$ be the measurement (i.e., evidence).

\[
P(c| x) = \frac{P(x|c)p(c)}{P(x)}
\]

- **class-conditional probability** (a.k.a. likelihood)
- **prior probability**
- **posterior probability**
- **unconditional probability** (a.k.a. marginal likelihood)
Example: 2D Bayes Classifier

17 samples
15 samples

These could be (g,b) pixel value of an image patch with grass

Given a (g,b) pixel value from a new patch is it more likely to be grass or sky?

These could be (g,b) pixel value of an image patch with sky

Image Credit: Ioannis (Yannis) Gkioulekas (CMU)
Example: 2D Bayes Classifier

- 17 samples
- 15 samples

\[ p(\text{blue}) = \frac{17}{17 + 15} \]

\[ p(\text{green}) = \frac{15}{17 + 15} \]

Image Credit: Ioannis (Yannis) Gkioulekas (CMU)
Example: 2D Bayes Classifier

- 17 samples
- 15 samples

\[ p(\text{blue}) = \frac{17}{17 + 15} \]

\[ p(\text{green}) = \frac{15}{17 + 15} \]

\[ p(\cdot | \text{green}) = \mathcal{N}(\mu_{\text{green}}, \Sigma_{\text{green}}) \]

\[ p(\cdot | \text{blue}) = \mathcal{N}(\mu_{\text{blue}}, \Sigma_{\text{blue}}) \]
**Example: 2D Bayes Classifier**

\[
p(green|\triangle) \propto \mathcal{N}(\triangle; \mu_{green}, \Sigma_{green})p(green)
\]

\[
p(blue|\triangle) \propto \mathcal{N}(\triangle; \mu_{blue}, \Sigma_{blue})p(blue)
\]

- 17 samples
- 15 samples

\[
p(blue) = \frac{17}{17 + 15}
\]

\[
p(green) = \frac{15}{17 + 15}
\]

\[
p(\cdot|green) = \mathcal{N}(\mu_{green}, \Sigma_{green})
\]

\[
p(\cdot|blue) = \mathcal{N}(\mu_{blue}, \Sigma_{blue})
\]
Loss Functions and Classifiers

Loss

- Some errors may be more expensive than others

**Example**: A fatal disease that is easily cured by a cheap medicine with no side-effects. Here, false positives in diagnosis are better than false negatives

- We discuss two class classification:
  \[ L(1 \rightarrow 2) \] is the loss caused by calling 1 a 2

**Total risk** of using classifier \( s \) is

\[
R(s) = \Pr\{1 \rightarrow 2 \mid \text{using } s\} \cdot L(1 \rightarrow 2) + \Pr\{2 \rightarrow 1 \mid \text{using } s\} \cdot L(2 \rightarrow 1)
\]
Two Class Classification

Generally, we should classify as 1 if the expected loss of classifying as 1 is less than for 2

Classify $x$ as

1 if $p(1|x) \cdot L(1 \rightarrow 2) > p(2|x) \cdot L(2 \rightarrow 1)$

2 if $p(1|x) \cdot L(1 \rightarrow 2) < p(2|x) \cdot L(2 \rightarrow 1)$

Decision boundary: points where the loss is the same for either class.
Training Error, Testing Error, and Overfitting

**Training error** is the error a classifier makes on the training set.

We want to minimize the **testing error** – the error the classifier makes on an unseen testing set.

Classifiers that have small training error may not necessarily have small testing error.

The phenomenon that causes testing error to be worse than training error is called **overfitting**.
Training Error, Testing Error, and **Overfitting**

**Underfitting**: model is too simple to represent all the relevant class characteristics
Training Error, Testing Error, and **Overfitting**

**Underfitting**: model is too simple to represent all the relevant class characteristics

**Overfitting**: model is too complex and fits irrelevant characteristics (noise) in the data
Training Error, Testing Error, and **Overfitting**

**Underfitting:** model is too simple to represent all the relevant class characteristics

**Overfitting:** model is too complex and fits irrelevant characteristics (noise) in the data

![Diagram showing training error, testing error, and overfitting](image)
Cross-Validation

We cannot reliably estimate the error rate of the classifier using the training set.

An alternative is to split some training data to form a validation set, then train the classifier on the rest of the data and evaluate on the validation set.

Try out what hyperparameters work best on test set.
Cross-Validation

We cannot reliably estimate the error rate of the classifier using the training set. An alternative is to split some training data to form a validation set, then train the classifier on the rest of the data and evaluate on the validation set.

Trying out what hyperparameters work best on test set:
Very bad idea. The test set is a proxy for the generalization performance! Use only VERY SPARINGLY, at the end.
We cannot reliably estimate the error rate of the classifier using the training set.

An alternative is to split some training data to form a validation set, then train the classifier on the rest of the data and evaluate on the validation set.
**Cross-Validation**

**Cross-validation** involves performing multiple splits and averaging the error over all splits.
Confusion Matrix

When evaluating a multi-class classifier, it may be useful to know how often certain classes are often misclassified as others.

A **confusion matrix** is a table whose (i,j)th entry is the frequency (or proportion) an item of true class i was labelled as j by the classifier.

Forsyth & Ponce (2nd ed.) Figure 15.3. Original credit: H. Zhang et al., 2006.
ROC curves plot trade-off between false positives and false negatives.

Forsyth & Ponce (2nd ed.) Figure 15.4

Figure from M. J. Jones and J. Rehg, “Statistical color models with application to skin detection,” Proc. CVPR, 1999, IEEE
Receiver Operating Characteristics (ROC)

What is a ROC curve for a perfect classifier?

ROC curves plot trade-off between false positives and false negatives.

Figure from M. J. Jones and J. Rehg, “Statistical color models with application to skin detection,” Proc. CVPR, 1999, IEEE
Receiver Operating Characteristics (ROC)

What is a ROC curve for a perfect classifier?

**ROC curves** plot trade-off between false positives and false negatives.

Forsyth & Ponce (2nd ed.) Figure 15.4

Figure from M. J. Jones and J. Rehg, “Statistical color models with application to skin detection,” Proc. CVPR, 1999, IEEE
Classifier Strategies

Classification strategies fall under two broad types: parametric and non-parametric.
Classifier Strategies

Classification strategies fall under two broad types: **parametric** and **non-parametric**.

Parametric classifiers are **model driven**. The parameters of the model are learned from training examples. New data points are classified by the learned model.

— fast, compact
— flexibility and accuracy depend on model assumptions
Classifier Strategies

Classification strategies fall under two broad types: **parametric** and **non-parametric**.

Parametric classifiers are **model driven**. The parameters of the model are learned from training examples. New data points are classified by the learned model.
- fast, compact
- flexibility and accuracy depend on model assumptions

Non-parametric classifiers are **data driven**. New data points are classified by comparing to the training examples directly. "The data is the model".
- slow
- highly flexible decision boundaries
Nearest Neighbor Classifier

Given a new data point, assign the label of nearest training example in feature space.

Image Credit: Ioannis (Yannis) Gkioulekas (CMU)
**Nearest Neighbor Classifier**

Given a new data point, assign the label of nearest training example in feature space.

*Image Credit: Ioannis (Yannis) Gkioulkas (CMU)*
**k-Nearest Neighbor (kNN) Classifier**

We can gain some robustness to noise by voting over multiple neighbours.

Given a new data point, find the k nearest training examples. Assign the label by majority vote.

Simple method that works well if the distance measure correctly weights the various dimensions.

For large data sets, as k increases kNN approaches optimality in terms of minimizing probability of error.
k-Nearest Neighbor (kNN) Classifier

kNN decision boundaries respond to local clusters where one class dominates

Figure credit: Hastie, Tibshirani & Friedman (2nd ed.)
Support Vector Machines (SVM)

Idea: Try to obtain the decision boundary directly

The decision boundary is parameterized as a separating hyperplane in feature space.
— e.g. a separating line in 2D

We choose the hyperplane that is as far as possible from each class - that maximizes the distance to the closest point from either class.
**Linear Classifier**

Defines a score function:

\[ f(x_i, W, b) = W x_i + b \]
Linear Classifier

Example with an image with 4 pixels, and 3 classes (cat/dog/ship)

Image Credit: Ioannis (Yannis) Gkioulkekas (CMU)
Support Vector Machines (SVM)

Learn the decision boundary

Image Credit: Ioannis (Yannis) Gkioulekas (CMU)
Support Vector Machines (SVM)

What’s the best \( w \) ?
Support Vector Machines (SVM)

What’s the best $w$?
Support Vector Machines (SVM)

What’s the best $w$?
Support Vector Machines (SVM)

What’s the best $w$?
Support Vector Machines (SVM)

What’s the best \( w \) ?

Image Credit: Ioannis (Yannis) Gkioulekas (CMU)
Support Vector Machines (SVM)

What’s the best $w$?

Intuitively, the line that is the farthest from all interior points

Image Credit: Ioannis (Yannis) Gkioulkas (CMU)
Support Vector Machines (SVM)

What’s the best $w$?

Want a hyperplane that is far away from ‘inner points’
Support Vector Machines (SVM)

Find hyperplane \( \mathbf{w} \) such that …

\[
\begin{align*}
\mathbf{w} \cdot \mathbf{x} + b &= 1 \\
\mathbf{w} \cdot \mathbf{x} + b &= 0 \\
\mathbf{w} \cdot \mathbf{x} + b &= -1
\end{align*}
\]

the gap between parallel hyperplanes \( \frac{2}{\|\mathbf{w}\|} \) is maximized

Slide Credit: Ioannis (Yannis) Gkioulkas (CMU)
Support Vector Machines (SVM)
Example: Pedestrian Detection with SVM

Figure credit: Papageorgiou, Oren, and Poggio, 1998
Summary

A classifier accepts as input a set of features and outputs (predicts) a class label.

Classifiers need to take into account “loss” associated with each kind of classification error.

A Receiver Operating Characteristic (ROC) curve plots the trade-off between false negatives and false positives.

**Parametric** classifiers are model driven. The parameters of the model are learned from training examples.
- e.g. support vector machine, decision tree.

**Non-parametric** classifiers are data driven. New data points are classified by comparing to the training examples directly.
- e.g. k-nearest neighbour.