1 Bayes Classifier

In the lecture we saw that the Bayes classifier is

$$c^*(x) := \arg\max_{y \in \mathcal{Y}} p(y|x).$$

(1)

a) Which of these decision functions is equivalent to $c^*$?

- $c_1(x) := \arg\max_y p(x)$
- $c_2(x) := \arg\max_y p(y)$
- $c_3(x) := \arg\max_y p(x, y)$
- $c_4(x) := \arg\max_y p(x|y)$

For $\mathcal{Y} = \{-1, +1\}$, we can express the Bayes classifier as

$$c^*(x) = \text{sign}[\log \frac{p(+1|x)}{p(-1|x)}].$$

b) Which of the following expressions are equivalent to $c^*$?

- $c_5(x) := \text{sign}[\log p(x, +1) - \log p(x, -1)]$
- $c_6(x) := \text{sign}[\log p(+1|x) + \log p(-1|x)]$
- $c_7(x) := \text{sign}[\log p(+1|x) - \log p(-1|x)]$
- $c_8(x) := \text{sign}[\log p(x, +1) - \log p(x, -1)]$
- $c_9(x) := \text{sign}[p(+1|x) - p(-1|x)]$
- $c_{10}(x) := \text{sign}\left[\frac{p(x, +1)}{p(x, -1)} - 1\right]$
- $c_{11}(x) := \text{sign}\left[\frac{\log p(+1|x)}{\log p(-1|x)} - 1\right]$
- $c_{12}(x) := \text{sign}[\log p(x, +1) + \log p(x, -1)]$

2 Risk Minimization

In medical diagnosis the loss function can be highly asymmetric. Let $y \in \{\text{yes, no}\}$ be the decision is whether a patient has cancer.

Assume $\ell(\text{no, no}) = \ell(\text{yes, yes}) = 0$ (correct diagnosis), $\ell(\text{no, yes}) = 1$ (the patient is upset until the further tests come in), $\ell(\text{yes, no}) = 1000$ (the cancer remains untreated and the patient might die).

For $\mathcal{X} = \mathbb{R}^+$ (e.g. the intensity of a certain region on the X-ray), assume a probability density:

$$p(x, y) = \begin{cases} 
  e^{-1.1x} & \text{for } y = \text{no}, \\
  e^{-11x} & \text{for } y = \text{yes},
\end{cases}$$

(2)

a) Show that $p(x, y)$ is a valid probability density.

b) Compute $p(x)$, $p(y)$, $p(y|x)$ and $p(x|y)$.

c) Compute the Bayes classifier $c^*$.

d) Compute the classifier $c^*_\ell$ that minimizes the $\ell$-risk.

e) Compute the expected risks $\mathcal{R}(c^*)$ and $\mathcal{R}(c^*_\ell)$.

f) What’s the simplest classifier you can come up with that has smaller $\ell$-risk than $c^*$?

3 Naive Bayes

a) Use the dating data from exercise sheet 1 to construct a discrete Naive Bayes classifier. Determine the training and the test error rate.

b) Does the dating data fulfill the implicit NB assumption that attributes are independent if conditioned on the labels?
4 Maximum Likelihood Estimate

Let \( \hat{p}(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) \) be a one-dimensional Gaussian model with parameters \( \mu \) and \( \sigma^2 \).

a) Show that for a set of i.i.d. examples \( x^1, \ldots, x^n \) of \( p(x) \),

\[
\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x^i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x^i - \mu)^2
\]

provide the maximum likelihood estimates, i.e. these parameters maximize \( \hat{p}(x^1, \ldots, x^n; \mu, \sigma^2) \).

Tip 1) maximizing \( \hat{p} \) is equivalent to maximizing \( \log \hat{p} \),

Tip 2) remember to use the i.i.d. assumption,

Tip 3) you can do the calculations for \( \mu \) and \( \sigma \) separately.

5 Gaussian Discriminant Analysis

Gaussian Discriminant Analysis (GDA) is an easy-to-compute method for generative probabilistic classification. For a training set \( \mathcal{D} = \{(x^1, y^1), \ldots, (x^n, y^n)\} \) set

\[
\mu := \frac{1}{n} \sum_{i=1}^{n} x^i, \quad \Sigma := \frac{1}{n} \sum_{i=1}^{n} (x^i - \mu)(x^i - \mu)^\top, \quad \mu_y := \frac{1}{|\{i : y^i = y\}|} \sum_{\{i : y^i = y\}} x^i, \quad \text{for } y \in \mathcal{Y},
\]

and define

\[
p(x|y) = \frac{1}{\sqrt{2\pi \det \Sigma}} \exp\left(-\frac{1}{2} (x - \mu_y)^\top \Sigma^{-1} (x - \mu_y)\right)
\]

a) Show for binary classification tasks: GDA leads to a linear decision rule, regardless of what \( p(y) \) is.

b) GDA is often used when there are only few examples available for each class. Can you imagine why?

6 Practical Experiments*

Implement the described Naive Bayes, Gaussian Discriminant Analysis and Logistic Regression classifiers. To solve the LogReg optimization, use any existing function optimizer, e.g. based on conjugate gradient or BFGS. Many language have this built in, otherwise, download a version from a web resource.

a) Train and test the classifier on the data from the previous exercise sheet.

7 Entertainment*

Watch the following video of "Watson", a machine learning system by IBM that plays Jeopardy!

http://www.youtube.com/watch?v=rxU1Pg-80as (or search for "Jeopardy! The IBM Challenge").

a) Discuss: How are Watson’s mistakes different from human mistakes? Why is that so?

b) Internally, Watson uses a probabilistic model similar to Logistic Regression (watch the outputs during the video). Why did the designers choose a discriminative probabilistic model instead of a generative one, or a directly learned decision function?