

Lecture 5

Probabilistic classifiers

Machine Learning
Andrey Filchenkov

04.10.2019

Lecture plan

- Bayesian classification
 - Non-parametric density recovery
 - Parametric density recovery
 - Logistic regression
-
- The presentation is prepared with materials of the K.V. Vorontsov's course "Machine Learning".
 - Slides are available online:
bit.ly/2lynu3L

Lecture plan

- Bayesian classification
- Non-parametric density recovery
- Parametric density recovery
- Logistic regression

Problem

An illness is spread among 1% of population. This illness test returns true answers in 95% of cases.

Someone receives a positive result.

What is the probability, he actually suffers the illness?

Problem: options

An illness is spread among 1% of population. This illness test returns true answers in 95% of cases. Someone receives a positive result. What is the probability, he actually suffers the illness?

$$97,5\% \leq x \leq 100\%$$

$$95\% \leq x < 97,5\%$$

$$92\% \leq x < 95\%$$

$$81\% \leq x < 92\%$$

$$70\% \leq x < 81\%$$

$$55\% \leq x < 70\%$$

$$30\% \leq x < 55\%$$

$$x < 30\%$$

Problem: answer

An illness is spread among 1% of population. This illness test returns true answers in 95% of cases. Someone receives a positive result. What is the probability, he actually suffers the illness?

$$\begin{aligned} \Pr(d = 1|t = 1) &= \\ &= \frac{\Pr(t = 1|d = 1) \Pr(d = 1)}{\Pr(t = 1|d = 1) \Pr(d = 1) + \Pr(t = 1|d = 0) \Pr(d = 0)} = \\ &= \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.05 \times 0.99} = \mathbf{0.16}. \end{aligned}$$

Probabilistic classification problem

Instead of an unknown target function $y^*(x)$, we will think about an unknown probability distribution on $X \times Y$ with a density $p(x, y)$.

Simple or independent identically distributed (i.i.d.) sample is a sample, which contains ℓ random independent observations $T^\ell = \{(x_i, y_i)\}_{i=1}^\ell$.

Now we have families of distributions $\{\varphi(x, y, \theta) | \theta \in \Theta\}$ instead of algorithm models.

Problem: find an algorithm, which minimizes probability of error.

Problem statement

$a: X \rightarrow Y$ splits X on non-overlapping domains A_y :

$$A_y = \{x \in X | a(x) = y\}.$$

An error occurs when object x with label y is classified as belonging to A_s , $s \neq y$.

Error probability: $\Pr(A_s, y) = \int_{A_s} p(x, y) dx$.

Error loss: $\lambda_{ys} \geq 0$, for all $(y, s) \in Y \times Y$.

Usually $\lambda_{yy} = 0$, $\lambda_y = \lambda_{ys} = \lambda_{yt} \ \forall s, t \in Y, s \neq y, t \neq y$.

Mean risk of a :

$$R(a) = \sum_{y \in Y} \sum_{s \in Y} \lambda_{ys} \Pr(A_s, y).$$

The main equation

$$p(X, Y) = p(x) \Pr(y|x) = \Pr(y) p(x|y)$$

$\Pr(y)$ is **priory probability** of class y .

$p(x|y)$ is **likelihood** of class y

$\Pr(y|x)$ is **posterior probability** of class y .

Two problems

First problem: **probability density recovering**

Given: $T^\ell = \{(x_i, y_i)\}_{i=1}^\ell$.

Problem: find empirical estimates $\widehat{\Pr}(y)$ and $\hat{p}(x|y)$, $y \in Y$.

Second problem: **mean risk minimization**

Given:

- prior probabilities $\Pr(y)$,
- likelihood $p(x|y)$, $y \in Y$.

Problem: find classifier a which minimizes $R(a)$.

Maximum a posteriori probability

Let $\Pr(y)$ and $p(x|y)$ be known for all $y \in Y$.

$$p(x, y) = p(x) \Pr(y|x) = \Pr(y) p(x|y).$$

Main idea: choose a class, such that the observation of this object belonging to this class is the most probable.

Maximum a posteriori probability (MAP):

$$a(x) = \operatorname{argmax}_{y \in Y} \Pr(y|x) = \operatorname{argmax}_{y \in Y} \Pr(y) p(x|y).$$

Optimal Bayesian classifier

Theorem

If $\Pr(y)$ and $p(x|y)$ are known then the minimal mean risk $R(a)$ is achieved by Bayesian classifier

$$a_{OB}(x) = \operatorname{argmin}_{s \in Y} \sum_{y \in Y} \lambda_{ys} \Pr(y) p(x|y).$$

If $\lambda_{yy} = 0, \lambda_y = \lambda_{ys} = \lambda_{yt} \ \forall s, t \in Y, s \neq y, t \neq y$

$$a_{OB}(x) = \operatorname{agrmin}_{y \in Y} \lambda_y \Pr(y) p(x|y).$$

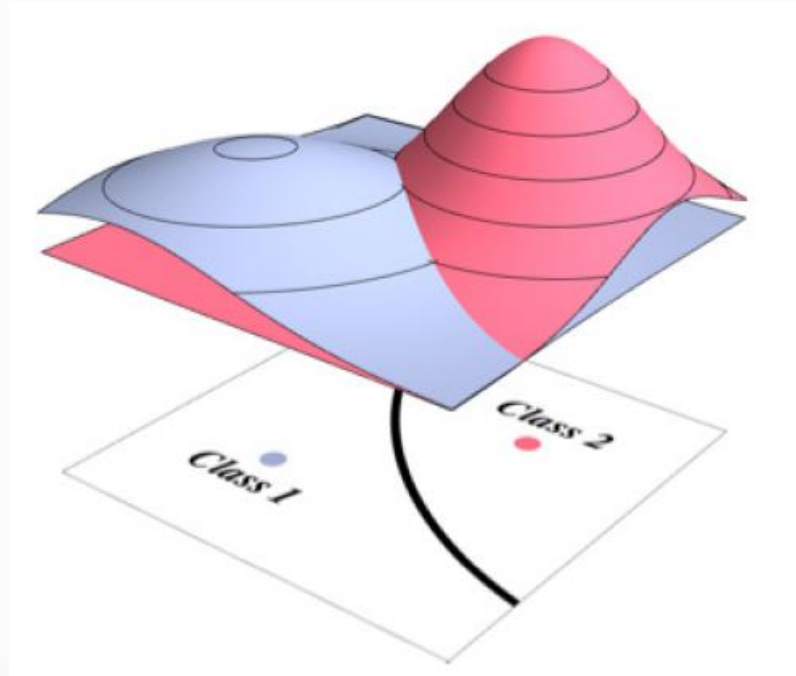
Classifier $a_{OB}(x)$ is **optimal Bayesian classifier**.

Bayesian risk is minimal value of $R(a)$.

Separating surface

Separating surface for classes a and b is locus of $x \in X$, such that the maximum of the Bayesian decision rule is achieved both for $y = s$ and $y = t$:

$$\lambda_a \Pr(a) p(x|a) = \lambda_s \Pr(b) p(x|b).$$



Lecture plan

- Bayesian classification
- **Non-parametric density recovery**
- Parametric density recovery
- Logistic regression

Two subproblems

The problem is to estimate prior and posterior probabilities for each class:

$$\begin{aligned}\widehat{\text{Pr}}(y) &=? \\ \hat{p}(x|y) &=?\end{aligned}$$

The first subproblem can be solved easily:

$$\widehat{\text{Pr}}(y) = \frac{|X_y|}{\ell}, \quad X_y = \{x_i, y_i \in T^\ell, y_i = y\}.$$

The second one is much more complex.

Loosing class

We can solve

$$\hat{p}(x|y) = ?$$

for each class independently.

Instead of recovering $\hat{p}(x|y)$, we will recover $\hat{p}(x)$ with $T^m = \left((x_{(1)}, s), \dots, (x_{(m)}, s) \right)$ for each $s \in Y$.

One-dimensional case

If $\Pr([a, b])$ is a probabilistic measure on $[a, b]$, then

$$p(x) = \lim_{h \rightarrow 0} \frac{1}{2h} \Pr([x - h, x + h]).$$

Empirical density estimation with window of a width h

$$\widehat{p}_h(x) = \frac{1}{2mh} \sum_{i=1}^m [|x - x_i| < h].$$

Parzen-Rosenblatt window

Empirical density estimation with window of a width h :

$$\widehat{p}_h(x) = \frac{1}{2hm} \sum_{i=1}^m \left[\frac{x - x_i}{h} < 1 \right].$$

Parzen-Rosenblatt estimation for a window with width h :

$$\widehat{p}_h(x) = \frac{1}{hm} \sum_{i=1}^m K\left(\frac{x - x_i}{h}\right),$$

where $K(r)$ is a kernel function.

$\widehat{p}_h(x)$ converges to $p(x)$.

Generalization to multidimensional case

1. If objects are described with n numeric features $f_j: X \rightarrow \mathbb{R}, j = 1, \dots, n$,

$$\widehat{p}_h(x) = \frac{1}{m} \sum_{i=1}^m \prod_{j=1}^n \frac{1}{h_j} K \left(\frac{f_j(x) - f_j(x_i)}{h_j} \right).$$

2. If X is a (metric) space with a distance $\rho(x, x')$:

$$\widehat{p}_h(x) = \frac{1}{mV(h)} \sum_{i=1}^m K \left(\frac{\rho(x, x_i)}{h} \right),$$

where $V(h) = \int_X K \left(\frac{\rho(x, x_i)}{h} \right) dx$ is normalizing factor.

Multidimensional Parzen window

Estimate $\widehat{p}_h(x)$ with

$$\widehat{p}_h(x) = \frac{1}{mV(h)} \sum_{i=1}^m K\left(\frac{\rho(x, x_i)}{h}\right),$$

Parzen window:

$$a(x; T^\ell, h) = \arg \max_{y \in Y} \lambda_y \Pr(y) \ell_y^{-1} \sum_{i: y_i = y} K\left(\frac{\rho(x, x_i)}{h}\right).$$

$\Gamma_y(x) = \lambda_y \Pr(y) \ell_y^{-1} \sum_{i: y_i = y} K\left(\frac{\rho(x, x_i)}{h}\right)$ is a closeness to class.

Naïve Bayesian classifier

Hypothesis (naïve): features are independent random variables with probability densities $p_j(\xi|y)$, $y \in Y$, $j = 1, \dots, n$.

Then classes likelihoods can be represented as:

$$p(x|y) = p_1(\xi_1|y) \cdot \dots \cdot p_n(\xi_n|y), \quad x = (\xi_1, \dots, \xi_n), y \in Y.$$

Naïve Bayesian classifier:

$$a(x) = \operatorname{argmax}_{y \in Y} \left(\ln \lambda_y \widehat{\Pr}(y) + \sum_{j=1}^n \ln \widehat{p}_j(\xi_j|y) \right).$$

Lecture plan

- Bayesian classification
- Non-parametric density recovery
- **Parametric density recovery**
- Logistic regression

Parametrical notation

Joint probability density for sample:

$$p(T^\ell) = p((x_1, y_1), \dots, (x_\ell, y_\ell)) = \prod_{i=1}^{\ell} p(x_i, y_i).$$

Likelihood:

$$L(\theta, T^\ell) = \prod_{i=1}^{\ell} \varphi(x_i, y_i, \theta).$$

MAP:

$$a_\theta(x) = \operatorname{argmax}_y \varphi(x, y, \theta).$$

Relation with empirical risk

Find logarithm:

$$-\ln L(\theta, T^\ell) = -\sum_{i=1}^{\ell} \ln \varphi(x_i, y_i, \theta) \rightarrow \min_{\theta} .$$

Define loss function:

$$L(a_{\theta}, x) = -\ell \ln \varphi(x, y, \theta) .$$

Then empirical risk minimization problem is:

$$\begin{aligned} Q(a_{\theta}, T^\ell) &= \frac{1}{\ell} \sum_{i=1}^{\ell} L(a_{\theta}, x) = \\ &= -\frac{1}{\ell} \sum_{i=1}^{\ell} \ell \ln \varphi(x_i, y_i, \theta) = -\sum_{i=1}^{\ell} \ln \varphi(x_i, y_i, \theta) \rightarrow \min_{\theta} . \end{aligned}$$

Maximum likelihood

Principle of **maximum likelihood**:

$$L(\theta; X^m) = \sum_{i=1}^m \ln \varphi(x_i; \theta) \rightarrow \max_{\theta},$$

Optimum for θ is achieved in a point in which the derivate value is zero.

Principle of maximum weighted likelihood:

$$L(\theta; X^m, W^m) = \sum_{i=1}^m w_i \ln \varphi(x_i; \theta) \rightarrow \max_{\theta},$$

where $W^m = \{w_1, \dots, w_m\}$ is a vector of object weights.

Maximum joint likelihood principle

$$Q(a_{\theta}, T^{\ell}) = - \sum_{i=1}^{\ell} \ln \varphi(x_i, y_i, \theta) \rightarrow \min_{\theta} .$$

$$\varphi(x_i, y_i, \theta) = p(x_i, y_i | w) p(w, \gamma),$$

$p(x_i, y_i | w)$ is a probabilistic data model, $p(w, \gamma)$ is prior distribution of model parameters, γ is hyper-parameter.

Maximum joint likelihood principle:

$$\sum_{i=1}^{\ell} \ln p(x_i, y_i | w) + \ln p(w, \gamma) \rightarrow \max_{w, \gamma} .$$

Quadratic penalty conditions

Let $w \in \mathbb{R}^n$ is described with n -dimensional Gaussian distribution:

$$p(w; \sigma) = \frac{1}{(2\pi\sigma)^{n/2}} \exp\left(-\frac{\|w\|^2}{2\sigma}\right),$$

(weights are independent, their expectations are equal to zeros, their variances are the same and equal to σ).

It leads to quadratic penalty:

$$-\ln p(w; \sigma) = \frac{1}{2\sigma} \|w\|^2 + \text{const}(w).$$

Lecture plan

- Bayesian classification
- Non-parametric density recovery
- Parametric density recovery
- Logistic regression

Linear classifiers (reminder)

Constraint: $Y = \{-1, +1\} = \{y_{-1}, y_{+1}\}$

Linear classifier:

$$a_w(x, T^\ell) = \text{sign} \left(\sum_{i=1}^n w_i f_i(x) - w_0 \right).$$

where $w_1, \dots, w_n \in \mathbb{R}$ are features weights.

$$a_w(x, T^\ell) = \text{sign}(\langle w, x \rangle).$$

Logistic regression

Idea: Instead of returning a binary answer, we can return more information, for instance, probability of belonging to a class.

In logistic regression, σ function is used instead of sign:

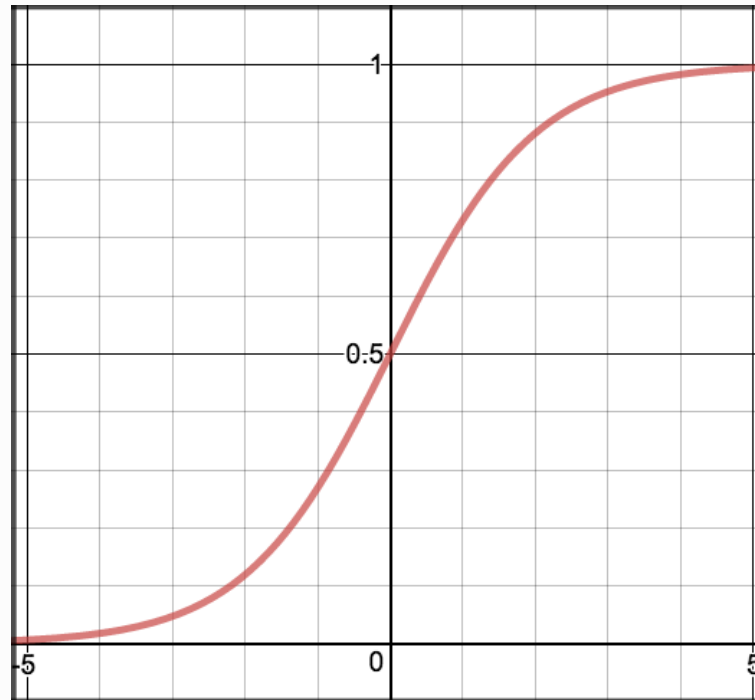
$$a_w(x, T^\ell) = \sigma(\langle w, x \rangle),$$

where

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Sigmoid (logistic) function

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$



Bayesian classification

A distribution $p(x, y)$ on object-answers space.

Simple sample of size ℓ

$$T^\ell = \{(x_i, y_i)\}_{i=1}^\ell.$$

Bayesian classifier:

$$a_{OB}(x) = \operatorname{argmax}_{y \in Y} \lambda_y \operatorname{Pr}(y) p(x|y),$$

where λ_y is losses for class y .

Linear Bayesian classifiers

$$Q(a_\theta, T^\ell) = \frac{1}{\ell} \sum_1^\ell L(a_\theta, x) = - \sum_{i=1}^\ell \ln \varphi(x_i, y_i, \theta) \rightarrow \min_\theta .$$

Bayesian classifier for two classes:

$$\begin{aligned} a(x) &= \text{sign}(\lambda_+ \Pr(y_+|x) - \lambda_- \Pr(y_- |x)) = \\ &= \text{sign} \left(\frac{p(x|y_+)}{p(x|y_-)} - \frac{\lambda_- \Pr(y_-)}{\lambda_+ \Pr(y_+)} \right) . \end{aligned}$$

Separating surface

$$\lambda_+ \Pr(y_+) p(x|y_+) = \lambda_- \Pr(y_-) p(x|y_-)$$

is linear.

Key hypothesis

Key hypothesis: classes are defined with n -dimensional overdispersed exponential densities:

$$p(x|y) = \exp \left(c_y(\delta) \langle \theta_y, x \rangle + b_y(\delta, \theta_y) + d(x, \delta) \right),$$

where $\theta_y \in \mathbb{R}^m$ is **shift** parameter,

δ is **dispersion** parameter;

b_y, c_y, d are some numeric functions.

Overdispersed exponential distribution family includes: uniform, normal, hypergeometric, Poisson, binominal, Γ -distribution and other.

Example: Gaussian

Let $\theta = \Sigma^{-1}\mu$; $\delta = \Sigma$.

Then

$$\begin{aligned}\mathcal{N}(x; \mu, \Sigma) &= \frac{e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}}{\sqrt{(2\pi)^n \det \Sigma}} = \\ &= \exp\left((\mu^\top \Sigma^{-1}x) - \left(\frac{1}{2}\mu^\top \Sigma^{-1}\Sigma\Sigma^{-1}\mu\right)\right)\end{aligned}$$

The main theorem

Theorem:

If p_y are overdispersed exponential distributions and $f_0(x) = \text{const}$, then

1) Bayesian classifier

$$a(x) = \text{sign} \left(\frac{p(x|y_+)}{p(x|y_-)} - \frac{\lambda_- \text{Pr}(y_-)}{\lambda_+ \text{Pr}(y_+)} \right)$$

is linear: $a(x) = \text{sign}(\langle w, x \rangle - w_0)$, $w_0 = \ln \frac{\lambda_-}{\lambda_+}$;

2) posterior probabilities of classes are:

$$\text{Pr}(y|x) = \sigma(\langle w, x \rangle y).$$

Logarithmic loss function

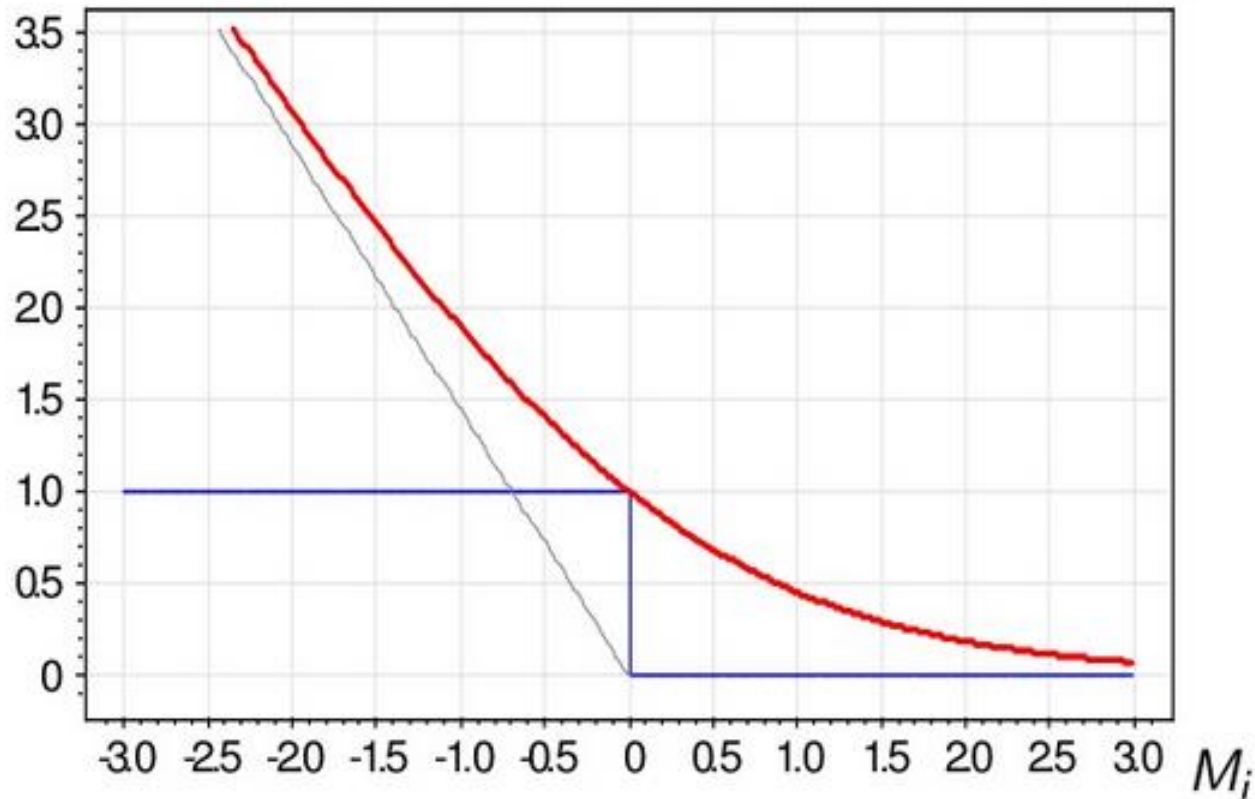
$$\widetilde{Q}_w(a, T^\ell) = \sum_i^\ell L(a, x_i) = \sum_i^\ell \ln p(x_i, y_i; w)$$

$$p(x, y; w) = \Pr(y|x)p(x) = \sigma(\langle w, x \rangle y) \text{const}(w)$$

$$\widetilde{Q}_w(a, T^\ell) = \sum_i^\ell \ln(1 + \exp(-\langle w, x \rangle y)) \rightarrow \min_w.$$

That is logarithmic loss function.

Logarithmic loss function plot



Gradient descent

Derivative:

$$\sigma'(s) = \sigma(s)\sigma(-s).$$

Gradient:

$$\mu \nabla \tilde{Q}(w^{[k]}) = - \sum_i^{\ell} y_i x_i \sigma(-M_i(w)).$$

Gradient descent step:

$$w^{[k+1]} = w^{[k]} - \mu y_i x_i \sigma(-M_i(w^{[k]})).$$