Statistical Machine Learning (W4400)
Spring 2016
https://courseworks.columbia.edu

Si Kai Lee
sl3950

Homework 1

Due: Tuesday 09 February 2016

Students are encouraged to work together, but homework write-ups must be done individually and must be entirely the author's own work. Homework is due at the **beginning** of the class for which it is due. **Late homework will not be accepted under any circumstances.** To receive full credit, students must thoroughly explain how they arrived at their solutions and include the following information on their homeworks: name, UNI, homework number (e.g., HW03), and class (STAT W4400). All homework must be turned in online through Courseworks in PDF format, have a .pdf extension (not zip or other archive!), and be less than 4MB. If programming is part of the assignment, the code must be turned in in one or more .R files. Homeworks not adhering to these requirements will receive no credit. For your convenience (not required), a tex template for producing nice PDF files can be found on courseworks.

1. **Naive Bayes** (20 points)

   Consider a classification problem with training data $\{(\tilde{\mathbf{x}}_1, \tilde{y}_1), \ldots, (\tilde{\mathbf{x}}_n, \tilde{y}_n)\}$ and three classes $\mathcal{C}_1$, $\mathcal{C}_2$ and $\mathcal{C}_3$. The sample space is $\mathbb{R}^5$, so each data point is of the form $\mathbf{x} = (x^{(1)}, \ldots, x^{(5)})$. Suppose we have reason to believe that the distribution of each class is reasonably well-approximated by a spherical (unit-variance) Gaussian, i.e. the class-conditional distributions are $g(\mathbf{x}|\mu_k, \mathbb{I})$ for class $k \in \{1, 2, 3\}$.

   1. How is the Gaussian assumption translated into a naive Bayes classifier? Write out the full formula for the estimated class label $\hat{y}_{\text{new}} = f(\mathbf{x}_{\text{new}})$ for a newly observed data point $\mathbf{x}_{\text{new}}$.
      Hint: This equation should not contain the training data, only parameters estimated from the training data.
   2. How do you estimate the parameters of the model? Give the estimation equations for (a) the parameters of the class-conditional distributions and (b) the class prior $P(y = k)$ for each class $\mathcal{C}_k$.
   3. If our assumptions on the data source as described above are accurate, do you expect the naive Bayes classifier to perform well? Please explain your answer.

Solution:

1a. The Gaussian assumption implies sample independence as there is zero covariance between classes which enables $\Pr(\mathbf{x}|y)$ to be written in the form of a naive Bayes classifier $\prod_{j=1}^{d} \Pr(\mathbf{x}|y)$.

1b. $\hat{y}_{new} = \underset{y\in\{C_1,C_2,C_3\}}{\arg\max} \Pr(y|\mathbf{x}_{new}) = \underset{y\in\{C_1,C_2,C_3\}}{\arg\max} \Pr(\mathbf{x}_{new}|y)\Pr(y)$, $\Pr(y) = \frac{1}{(2\pi\sigma)^{\frac{1}{2}}} e^{-\frac{(x-\mu_k)^2}{2}}$

2a. For each class C_k, $\mu_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \mathbf{x}_k$

2b. $\Pr(y = k) = \frac{\text{\# observations in } C_k}{\text{\#observations}}$

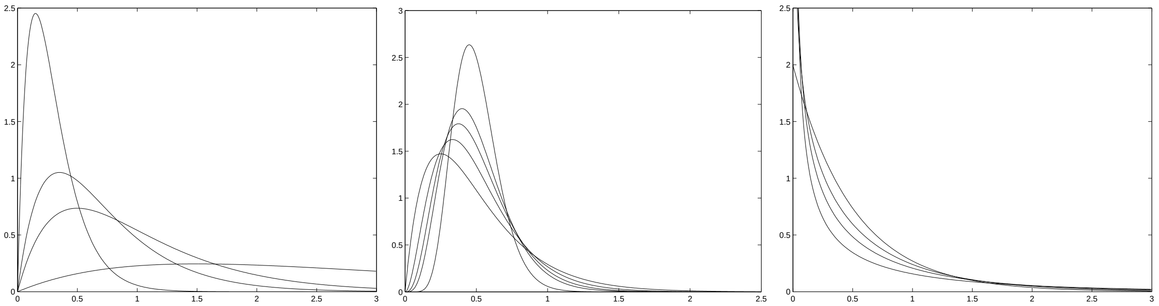

Figure 1: *Left:* The plot shows the density for different values of the location parameter ($\mu = 0.3, 0.5, 1.0, 3.0$), with the shape parameter fixed to $\nu = 2$. Since $\nu > 1$, the densities peak. As we increase μ, the peak moves to the right, and the curve flattens. *Middle:* For $\mu = 0.5$ fixed, we look at different values of the shape parameter ($\nu = 2, 3, 4, 5, 19$). Again, all the densities peak, with the peak shifting to the right as we increase ν. *Right:* If $\nu < 1$, the density turns into a monotonously decreasing function. The smaller the value of ν, the sharper the curve dips towards the origin.

3. Yes. Given that the samples in the training data are parameterised by an identity covariance matrix denoting zero covariance between classes, they fit the sample independence assumption made by the naive Bayes classifier. Hence the classifier would classify the samples well as its assumptions are in line with the nature of the dataset.

2. **Maximum Likelihood Estimation** (40 points)

   In this problem, we analytically derive maximum likelihood estimators for the parameters of an example model distribution, the gamma distribution.

   The gamma distribution is univariate (one-dimensional) and continuous. It is controlled by two parameters, the *location parameter* μ and the *shape parameter* ν. For a gamma-distributed random variable X, we write $X \sim \mathcal{G}(\mu, \nu)$. $\mathcal{G}$ is defined by the following density function:

$$p(x|\mu, \nu) := \left(\frac{\nu}{\mu}\right)^{\nu} \frac{x^{\nu-1}}{\Gamma(\nu)} \exp\left(-\frac{\nu x}{\mu}\right) ,$$

   where $x \geq 0$ and $\mu, \nu > 0$.[1] Whenever $\nu > 1$, the gamma density has a single peak, much like a Gaussian. Unlike the Gaussian, it is not symmetric. The first two moment statistics of the gamma distribution are given by

$$\mathsf{E}[X] = \mu \qquad \text{and} \qquad \mathsf{Var}[X] = \frac{\mu^2}{\nu} \tag{1}$$

   for $X \sim \mathcal{G}(\mu, \nu)$. The plots in Figure 1 should give you a rough idea of what the gamma density may look like and how different parameter values influence its behavior.

[1] The symbol Γ denotes the distribution's namesake, the *gamma function*, defined by

$$\Gamma(\nu) := \int_0^\infty e^{-t} t^{\nu-1} dt .$$

The gamma function is a generalization of the factorial to the real line: $\Gamma(n) = (n-1)!$ for all $n \in \mathbf{N}$. Fortunately, we will not have to make explicit use of the integral.

Questions:

- (8 points) Write the general analytic procedure to obtain the maximum likelihood estimator (including logarithmic transformation) in the form of a short algorithm or recipe. A few words are enough, but be precise: Write all important mathematical operations as formulae. Assume that data is given as an i. i. d. sample $x_1, \dots, x_n$. Denote the conditional density in question by $p\left(x|\theta\right)$, and the likelihood by $l\left(\theta\right)$. Make sure both symbols show up somewhere in your list, as well as a logarithm turning a product into a sum.
- (16 points) Derive the ML estimator for the location parameter μ, given data values $x_1, \dots, x_n$. Conventionally, an estimator for a parameter is denoted by adding a hat: $\hat{\mu}$. Considering the expressions in (1) for the mean and variance of the gamma distribution, and what you know about MLE for Gaussians, the result should not come as a surprise.
- (16 points) A quick look at the gamma density will tell you that things get more complicated for the shape parameter: ν appears inside the gamma function, and both inside and outside the exponential. Thus, instead of deriving a formula of the form $\hat{\nu} := \dots$, please show the following: Given an i. i. d. data sample $x_1, \dots, x_n$ and the value of μ, the ML estimator $\hat{\nu}$ for the gamma distribution shape parameter solves the equation

  $$\sum_{i=1}^{n} \left(\ln\left(\frac{x_i \hat{\nu}}{\mu}\right) - \left(\frac{x_i}{\mu} - 1\right) - \phi\left(\hat{\nu}\right) \right) = 0 \; .$$

  The symbol ϕ is a shorthand notation for

  $$\phi\left(\nu\right) := \frac{\frac{\partial \Gamma(\nu)}{\partial \nu}}{\Gamma\left(\nu\right)} \; .$$

  In mathematics, ϕ is known as the *digamma function*.

Solution:

Part A.

1. Set the likelihood function as $L_n(\theta, \mathbf{x}) = \prod\limits_{i=1}^{n} \Pr(x|\theta)$
2. Take $\ln$ of L_n to get $\ell_n(\theta, \mathbf{x}) = \ln \prod\limits_{i=1}^{n} \Pr(x|\theta) = \sum\limits_{i=1}^{n} \ln \Pr(x|\theta)$
3. Differentiate ℓ_n w.r.t. θ to get $\frac{\partial \ell}{\partial \theta}$ (If θ is a vector, calculate $\nabla \ell_n$)
4. Solve $\frac{\partial \ell}{\partial \theta} = 0$ to get $\hat{\theta}$ (If θ is a vector, solve $\nabla \ell_n = 0$ to get $\hat{\theta}$)

Part B.

Get $\frac{\partial \ell}{\partial \mu} = \frac{\partial}{\partial \mu}(n\nu \ln \nu - n\nu \ln \mu + (\nu - 1)\ln(\prod\limits_{i=1}^{n} x) - n\ln(\Gamma(\nu)) - \frac{\nu \sum\limits_{i=1}^{n} x_i}{\mu}) = -\frac{n\nu}{\mu} + \frac{\nu \sum\limits_{i=1}^{n} x_i}{\mu^2}$

Equate $-\frac{n\nu}{\mu} + \frac{\nu \sum\limits_{i=1}^{n} x_i}{\mu^2}$ to 0 to get $\frac{n\nu}{\mu} = \frac{\nu \sum\limits_{i=1}^{n} x_i}{\mu^2}$ which provides the estimator for μ, $\hat{\mu} = \frac{1}{n} \sum\limits_{i=1}^{n} x_i$

Part C.

Get $\frac{\partial \ell}{\partial \mu} = \frac{\partial \ell}{\partial \nu} = n \ln \nu + n - n \ln \mu + \sum\limits_{i=1}^{n} \ln x - n \frac{\partial}{\partial \nu} \ln(\Gamma(\nu)) - \frac{\sum\limits_{i=1}^{n} x_i}{\mu}$

Equate above to 0 and collect terms:

- $n \ln \nu - n \ln \mu + \sum\limits_{i=1}^{n} \ln x_i = \sum\limits_{i=1}^{n} \ln \frac{\nu}{\mu} + \sum\limits_{i=1}^{n} \ln x_i = \sum\limits_{i=1}^{n} \ln \frac{x_i \nu}{\mu}$
- $-\frac{\sum\limits_{i=1}^{n} x_i}{\mu} + n = -\sum\limits_{i=1}^{n} (\frac{x_i}{\mu} - 1)$
- $-n \frac{\partial}{\partial \nu} \ln(\Gamma(\nu)) = -n \frac{\frac{\partial}{\partial \nu}\Gamma(\nu)}{\Gamma(\nu)} = -\sum\limits_{i=1}^{n} \frac{\frac{\partial}{\partial \nu}\Gamma(\nu)}{\Gamma(\nu)} = -\sum\limits_{i=1}^{n} \phi(\nu)$

Combine the collected terms to obtain $\sum\limits_{i=1}^{n} (\ln \frac{x_i \nu}{\mu} - (\frac{x_i}{\mu} - 1) - \phi(\nu)) = 0$

3. **Bayes-Optimal Classifier** (30 points)

Consider a classification problem with K classes and with observations in $\mathbb{R}^d$. Now suppose we have access to the true joint density $p(\mathbf{x}, y)$ of the data $\mathbf{x}$ and the labels y. From $p(\mathbf{x}, y)$ we can derive the conditional probability $P(y|\mathbf{x})$, that is, the posterior probability of class y given observation $\mathbf{x}$.

In the lecture, we have introduced a classifier f_0 based on p, defined as

$$f_0(\mathbf{x}) := \arg\max_{y \in [K]} P(y|\mathbf{x}) ,$$

the *Bayes-optimal classifier*.

Homework question: Show that the Bayes-optimal classifier is the classifier which minimizes the probability of error, under all classifiers in the hypothesis class

$$\mathcal{H} := \{ f \colon \mathbb{R}^d \to [K] \,|\, f \text{ integrable } \} .$$

(If you are not familiar with the notion of an integrable function, just think of this as the set of all functions from $\mathbb{R}^d$ to the set $[K]$ of class labels.)

Hints:

- The probability of error is precisely the risk under zero-one loss.
- You can greatly simplify the problem by decomposing the risk $R(f)$ into conditional risks $R(f|\mathbf{x})$:

$$R(f|\mathbf{x}) := \sum_{y \in [K]} L^{\text{0-1}}(y, f(\mathbf{x})) P(y|\mathbf{x}) \qquad \text{and hence} \qquad R(f) = \int_{\mathbb{R}^d} R(f|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} .$$

If you can show that f_0 minimizes $R(f|\mathbf{x})$ for every $\mathbf{x} \in \mathbb{R}^d$, the result for $R(f)$ follows by monotonicity of the integral.

Solution:
We want to show f_0 minimises $R(f|\mathbf{x})$ for every $\mathbf{x}$ so we substitute f_0 into $R(f|\mathbf{x})$ which gives $R(f_0) = \sum_{y \in [K]} L^{0-1}(y, f_0(x)) \Pr(y|\mathbf{x})$.

The loss function, $L^{0-1}(y, f_0(x))$, is an indicator function which outputs 0 if $f_0(\mathbf{x}) = y$ and 1 if $f_0(\mathbf{x}) \neq y$. Since the support of $\Pr(y|\mathbf{x}) = [0, 1]$, therefore the support of the risk $R(f_0|\mathbf{x}) = L^{0-1}(y, f_0(x)) \Pr(y|\mathbf{x})$ of any $\mathbf{x}$ is $[0, 1]$.

Say f_0 outputs A, the class of observation $\mathbf{x}_i$ in which $\Pr(y|\mathbf{x}_i)$ is largest. Then we know that the probability $\mathbf{x}_i$ belonging to any class in the set of $[K] - A$ would be lower than $\Pr(A|\mathbf{x}_i)$, hence the risk of misclassifying $\mathbf{x}_i$ would be always less than $\Pr(A|\mathbf{x}_i)$ as the loss function would output 0 if the true label of $\mathbf{x}$ is A. By ensuring that the largest component of $\Pr(y|\mathbf{x}_i)$, $\max_{y \in [K]} \Pr(y|\mathbf{x}_i)$, is not present in $R(f_0|\mathbf{x})$, we minimise $R(f_0|\mathbf{x})$ as the total probability of the classifier misclassifying x_i is less than $\max_{y \in [K]} \Pr(y|\mathbf{x}_i)$ which is the best result that the classifier can achieve.

By extending this argument for every $\mathbf{x} \in \mathbb{R}^d$, we have shown that the Bayes-optimal classifier f_0 minimises $R(f|\mathbf{x})$ for every $\mathbf{x}$ and hence $R(f)$.

4. **Risk** (10 points)
The following questions all consider a binary classifier $f : \mathbb{R}^d \to \{-1, +1\}$.

(a) (1 point) To calculate the risk $R(f)$, what function(s) are required, and why? An acceptable answer can write down the form of risk $R(f)$ and describe the components.

Solution:
$R(f) = \sum_{y \in \{0,1\}} \int L(y, f(\mathbf{x})) \Pr(\mathbf{x}, y) d\mathbf{x}$. We need a loss function L (usually $0-1$ loss) to calculate the performance of the algorithm for each $\mathbf{x}$, the classifier $f(\mathbf{x})$ to classify the observed $\mathbf{x}$ and joint distribution of $\mathbf{x}$ and y, $Pr(\mathbf{x}, y)$, which determines the probability of the class and given observation $\mathbf{x}_i$.

(b) (1 point) Do most machine learning algorithms use risk $R(f)$ or empirical risk $\hat{R}_n(f)$, and why?

Solution:
Most algorithms use $\hat{R}_n(f)$ to approximate $R(f)$ as we do not know the underlying distribution of y and $\mathbf{x}$.

(c) (2 points) If the training data $\{(x_1, y_1), ..., (x_n, y_n)\}$ for a fixed classifier f are n iid draws from the true underlying distribution of the data, what is:

$$\lim_{n \to \infty} \left| R(f) - \hat{R}_n(f) \right|$$

Please make a simple argument; no proof is required. (Technical note: you may assume that $R(f)$ is well behaved such that questions of convergence are all appropriately satisfied).

Solution:
0. Using the Weak Law of Large Numbers, we assume that after a large number of i.i.d. draws from the true underlying distribution, the empirical distribution would reflect the true underlying distribution. Hence $\lim_{n\to\infty} \left| R(f) - \hat{R}_n(f) \right| = 0.$

(d) (2 points) Under the usual $0-1$ loss, what is the range of $R(f)$? With this answer, interpret $R(f)$ in words as a probability (one sentence will suffice).

Solution:
The range (or support) of $R(f) = [0,1]$. $R(f)$ is the expected probability of an $\mathbf{x}_i$ classified incorrectly.

(e) (2 points) Training procedure 1 chooses linear classifiers f^1 entirely at random. Now the risk $R(f^1)$ is a random variable (a function of the random variable f^1). What is $E\left(R\left(f^1\right)\right)$ under the 01 loss?

Solution:
The analytic solution is $\mathbb{E}[R(f^1)] = \mathbb{E}[\sum\limits_{y\in\{0,1\}} \int L(y, f(\mathbf{x}))\Pr(\mathbf{x}, y)d\mathbf{x}]$ which we can write as $= \sum\limits_{y\in\{0,1\}} \int_{R(f^1)} \int_x L(y, f(\mathbf{x}))\Pr(\mathbf{x}, y)d\mathbf{x}\Pr(R(f^1))d(R(f^1))$. However since we are not given $\Pr(R(f^1)) = \Pr(f^1)$, we can assume that the set of linear classifiers are symmetrical as they can be reflected with respect to some line. Hence for each classifier that classifies m points correctly there is a classifier that classifies $n-m$ points so $R(f^1)$ can be assumed to follow a Unif$(0,1)$ distribution. Hence $\mathbb{E}[R(f^1)]$ is the the average probability of points classified correctly which is $\frac{1}{2}$.

(f) (2 points) Training procedure 2 uses the Perceptron to choose a linear classifier f^2 according to a training set $\{(x_1, y_1), ..., (x_n, y_n)\}$ drawn iid from the true underlying distribution. By analogy to the previous part, you can consider that training procedure 2 chooses linear classifiers f^2 *better than* entirely at random. Do you expect $E\left(R\left(f^2\right)\right)$ to be larger or smaller than $E\left(R\left(f^1\right)\right)$, again under the same 01 loss?

Solution:
If the training set is linearly separable, $E\left(R\left(f^2\right)\right)$ would be smaller $E\left(R\left(f^1\right)\right)$ as the Perceptron algorithm would converge to output a classifier with minimum risk. If the training set is not linearly separable, the $E\left(R\left(f^2\right)\right)$ would not be guaranteed to be smaller than $E\left(R\left(f^1\right)\right)$.