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

John P. Cunningham
jpc2181
Ben Reddy, Phyllis Wan, Ashutosh Nanda
bmr2136, pw2348, an2655

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:

1. (8 points) The estimate of the class label of $\mathbf{x}$ is
$$\hat{y} = \arg\max_{j \leq 3} \Big(\prod_{d=1}^{5} g(x^{(d)} | \mu_j^{(d)}, 1) \Big) \hat{P}(y = j)$$
2. (8 points) The parameters of the class-conditional Gaussians are estimated by maximum likelihood, seperately for separate dimensions, as
$$\hat{\mu}_j^{(d)} := \frac{1}{|\mathcal{C}_j|} \sum_{i \in \mathcal{C}_j} x_i^{(d)} \ .$$

The class priors estimates are

$$\hat{P}(y = j) = \frac{|\mathcal{C}_j|}{n} .$$

3. (4 points) The naive Bayes classifier assumes independence between dimensions; since dimensions are indeed independent in spherical Gaussians, the naive Bayes classifier should be quite accurate.

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.

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 $x_1, \ldots, x_n$. Denote the conditional density in question by $p(x|\theta)$, and the likelihood by $l(\theta)$. 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 $x_1, \ldots, x_n$. 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 $\hat{\nu} := \ldots$, please show the

[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.

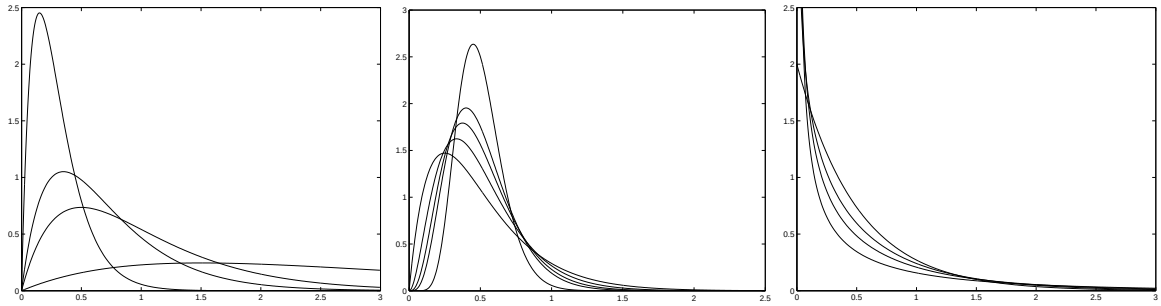

Figure 1: *Left:* The plot shows the density for different values of the location parameter ($\mu = 0.3, 0.5, 1.0, 3.0$), with the shape parameter fixed to $\nu = 2$. Since $\nu > 1$, the densities peak. As we increase μ, the peak moves to the right, and the curve flattens. *Middle:* For $\mu = 0.5$ 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.

following: Given an i. i. d. data sample $x_1, \ldots, x_n$ 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:

1 (8 points). A general recipe for deriving the MLE $\hat{\theta}$ of a distribution parameter θ involves the following steps:

1. Assume the data is given as an i. i. d. sample $X_1, X_2, \ldots, X_n \sim p(x; \theta)$.
2. Using the independence assumption, we rewrite $p(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} p(x_i; \theta) =: l(\theta)$.
3. The maximum of $l(\theta)$ is invariant under strictly monotonous transformations, therefore $\hat{\theta} = \arg\max_\theta \ln l(\theta) = \arg\max_\theta l(\theta)$.
4. We find the maximum by taking $\frac{\partial}{\partial \theta} \ln l(\theta) \stackrel{!}{=} 0$, while making sure that $\left[\frac{\partial^2}{\partial \theta^2} \ln l(\theta)\right]_{\theta=\hat{\theta}} < 0$.

2 (16 points). We derive the MLE $\hat{\mu}$ of the location parameter μ according to the recipe:

$$\begin{aligned}
\frac{\partial}{\partial \mu} \ln l(\mu, \nu) &= \frac{\partial}{\partial \mu} \ln \prod_{i=1}^{n} p(x_i; \mu, \nu) \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \mu} \ln \left[\left(\frac{\nu}{\mu} \right)^{\nu} \frac{x_i^{\nu-1}}{\Gamma(\nu)} \exp\left(-\frac{\nu x_i}{\mu} \right) \right] \\
&= \sum_{i=1}^{n} \left[\underbrace{\frac{\partial}{\partial \mu} \ln \left(\nu^{\nu} \frac{x_i^{\nu-1}}{\Gamma(\nu)} \right)}_{=0} + \frac{\partial}{\partial \mu} \ln \left(\mu^{-\nu} \exp\left(-\frac{\nu x_i}{\mu} \right) \right) \right] \\
&= \sum_{i=1}^{n} \left[-\frac{\partial}{\partial \mu} \nu \ln \mu + \frac{\partial}{\partial \mu} \left(-\frac{\nu x_i}{\mu} \right) \right] \\
&= \sum_{i=1}^{n} \left[-\frac{\nu}{\mu} + \frac{\nu x_i}{\mu^2} \right]
\end{aligned}$$

Now set $\frac{\partial}{\partial \theta} \ln l(\theta) \stackrel{!}{=} 0$ and rearrange

$$\begin{aligned}
\sum_{i=1}^{n} \left[-\frac{\nu}{\hat{\mu}} + \frac{\nu x_i}{\hat{\mu}^2} \right] &= 0 \\
\frac{\nu}{\hat{\mu}} \sum_{i=1}^{n} x_i &= n\nu \\
\sum_{i=1}^{n} x_i &= n\hat{\mu} \\
\hat{\mu} &= \frac{1}{n} \sum_{i=1}^{n} x_i \, .
\end{aligned}$$

With $\hat{\mu}$, we have a candidate for maximum, but the first derivative only tells us that it is a stationary point. To verify that it indeed maximizes the likelihood, we check that the second derivative is negative. The second derivative is:

$$\frac{\partial^2}{\partial \mu^2} \ln l(\mu, \nu) = \sum_{i=1}^{n} [\frac{\nu}{\mu^2} - \frac{2\nu x_i}{\mu^3}]$$

Substituting $\mu = \hat{\mu}$ gives:

$$\begin{aligned}
\sum_{i=1}^{n} [\frac{\nu}{\hat{\mu}^2} - \frac{2\nu x_i}{\hat{\mu}^3}] &= \frac{n\nu}{\hat{\mu}^2} - \frac{2\nu}{\hat{\mu}^3} \sum_{i=1}^{n} x_i \\
&= \frac{n\nu}{\bar{x}^2} - \frac{2\nu}{\bar{x}^3} n\bar{x} \\
&= -\frac{n\nu}{\bar{x}^2} \, ,
\end{aligned}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. By definition of the gamma distribution, the shape parameter ν is positive, which implies $(-n\nu)/\bar{x}^2 < 0$.

3 (16 points). For the gamma shape parameter ν:

$$\begin{aligned}
\frac{\partial}{\partial \nu} \ln l(\mu, \nu) &= \frac{\partial}{\partial \nu} \ln \prod_{i=1}^{n} p(x_i; \mu, \nu) \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \nu} \ln \left[\left(\frac{\nu}{\mu} \right)^{\nu} \frac{x_i^{\nu-1}}{\Gamma(\nu)} \exp\left(-\frac{\nu x_i}{\mu} \right) \right] \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \nu} \left[\nu \ln \frac{\nu}{\mu} + (\nu - 1) \ln x_i - \ln \Gamma(\nu) - \frac{\nu x_i}{\mu} \right] \\
&= \sum_{i=1}^{n} \left[\ln \frac{\nu}{\mu} + 1 + \ln x_i - \frac{\frac{\partial}{\partial \nu}\Gamma(\nu)}{\Gamma(\nu)} - \frac{x_i}{\mu} \right] \\
&= \sum_{i=1}^{n} \left[\ln \frac{x_i \nu}{\mu} - (\frac{x_i}{\mu} - 1) - \phi(\nu) \right] .
\end{aligned}$$

Setting $\left[\frac{\partial}{\partial \nu} \ln l(\mu, \nu) \right]_{\nu=\hat{\nu}} \stackrel{!}{=} 0$ leads to the desired result.

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

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

(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 $R(f|\mathbf{x})$:

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

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:

Since the probability of misclassification is precisely the risk under 0-1-loss (see Hint 1 on the homework sheet), we have to show that the classifier defined as

$$f_0(\mathbf{x}) := \arg\max_{y \in [K]} P(y|\mathbf{x}) \tag{2}$$

satisfies

$$f_0(\mathbf{x}) = \arg\min_{f \in \mathcal{H}} R(f) \ .$$

Recall that $R(f)$ was defined as

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

The integral makes working with this term cumbersome, but we can simplify it using Hint 2: Since

$$p(\mathbf{x}, y) = \int_{\mathbb{R}^d} P(y|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \ ,$$

we can rewrite $R(f)$ as

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

Still following Hint 2, it is now sufficient to minimize $R(f|\mathbf{x})$ "point-wise in $\mathbf{x}$". Put more simply: If we can find an f_0 which satisfies

$$f_0 = \arg\min_{f \in \mathcal{H}} R(f|\mathbf{x}) \qquad \text{for } \textbf{all } \mathbf{x} \in \mathbb{R}^d \ , \tag{3}$$

then this f_0 clearly minimizes $R(f)$ as well.

What we have to show now is that the classifier f_0 defined in (**??**) minimizes (**??**). By definition,

$$\begin{aligned} R(f|\mathbf{x}) &= \sum_{k \in [K]} P(y|\mathbf{x}) L^{0\text{-}1}(f(\mathbf{x}), y) \\ &= P(y = f(\mathbf{x})|\mathbf{x}) \cdot \underbrace{L^{0\text{-}1}(f(\mathbf{x}), f(\mathbf{x}))}_{=0} + \sum_{k \neq f(\mathbf{x})} P(k|\mathbf{x}) \cdot \underbrace{L^{0\text{-}1}(f(\mathbf{x}), k)}_{=1} \\ &= \sum_{k \neq f(\mathbf{x})} P(k|\mathbf{x}) \ . \end{aligned}$$

Since we know that $P(k|\mathbf{x})$ sums to 1 over all k, that means

$$R(f|\mathbf{x}) = 1 - P(f(\mathbf{x})|\mathbf{x}) \ .$$

But the definition in (**??**) says that f_0 maximizes $P(f(\mathbf{x})|\mathbf{x})$, and hence minimizes $1 - P(f(\mathbf{x})|\mathbf{x})$, at each $\mathbf{x}$. In summary,

$$f_0 = \arg\min_{f \in \mathcal{H}} R(f|\mathbf{x}) \qquad \text{for all } \mathbf{x} \in \mathbb{R}^d \ ,$$

and therefore

$$f_0 = \arg\min_{f \in \mathcal{H}} R(f) \ .$$

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: The distribution of the data and a loss function.

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

Solution: Empirical risk, due to uncertainty about the true distribution of the data.

(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.

(d) (2 points) Under the usual 01 loss, what is the range of $R(f)$? With this answer, interpret $R(f)$ in words as a probability (one sentence will suffice).

Solution: $[0, 1]$. $R(f)$ under the 01 loss is the probability that a given classifier f makes an error.

(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: 0.5

(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: Smaller. Risk under the 01 loss is an error measure, so the Perceptron will do better.