Notation

Kevin P. Murphy

Last updated October 25, 2006

1 General

- $x^* \in \arg\max_x f(x)$ means x^* is the value of x that maximizes the function f, i.e., $f(x^*) = \max_x f(x)$. Note that there may be multiple global maxima, in which case we break ties randomly.
- Indicator function: $I(e) = 1$ if event e is true, $I(e) = 0$ otherwise.
- Delta function: $\delta(x) = 1$ if $e = 0$ and $\delta(x) = 0$ otherwise.
- Sometimes probability mass functions (for discrete random variables) are written $P(X)$, and probability density functions (for continuous random variables) are written $p(X)$. We will use $p()$ for both.
- Usually we write random variables as capital letters and values of random variables as lower case, e.g., $p(X = x)$ is the probability X has value x. However, we do not follow this convention very closely.
- If X is distributed according to distribution f with parameters θ, i.e., $p(X) = f(X|\theta)$, then we write $X \sim f(\theta)$.
- We will often write probability distributions up to a constant of proportionality, $p(x|\theta) \propto f(x, \theta)$. This normalization constant is often denoted $1/Z(\theta)$, where Z is called the partition function.
- Vectors are usually column vectors. T denotes transpose, so x^T is a row vector. Sometimes we will write vectors in bold, e.g., $\mathbf{x}$, or as $\vec{x}$, but usually we will just write x. Matrices will usually be written as capital letters, X. However, using this convention we will cannot distinguish matrices from scalar (or vector) random variables. It should be clear from context.
- We use the following matlab notation: $1 : n$ denotes the sequence of integers $\{1, 2, \ldots, n\}$ and $X(i, j, k)$ is element i, j, k of some matrix, where i, j, k could each be a sequence of indices.

2 Data and variables

- X_{ni} is the i'th component/ feature / variable of data case n, for $i = 1 : D$, where D is the number of features/ variables, and $n = 1 : N$, where N is the number of training samples. (In general, D may depend on n if each example has a different size, but we will rarely write D_n.) If there is a single training/ test sample, we just write X_i for the i'th variable.
- $X = X(1 : N, 1 : D)$ is the design matrix. The n'th row is the n'th example X_n^T (since each example X_n is a column vector); the i'th column of X is all values of the i'th feature. We also write this as $\mathcal{D} = \{x_n\}_{n=1}^N$, which is more general notation, since it does not assume that all examples have the same number of features (e.g., a document may contain sentences of different lengths, so we would use $\mathcal{D}$ rather than X).
- In supervised learning problems, there is a distinguished output variable y_n, so $\mathcal{D} = \{x_n, y_n\}$. In classification, $y_n \in \{1, \ldots, C\}$, where C is the number of classes. In regression, $y_n \in \mathbb{R}$.

- If X_{ni} is a scalar, then $X_{ni} \in \mathbb{R}$ or $X_{ni} \in \mathbb{R}^+$. If it is a vector, then $X_{ni} \in \mathbb{R}^K$, where X_{nik} is the k'th component of the i'th variable for $k = 1 : K$, where K is the dimensionality of each variable. (In general, K may depend on i, but we rarely write K_i).
- If X_{ni} is binary, then $X_{ni} \in \{0, 1\}$. If X_{ni} is categorical, then $X_{ni} \in \{1, \ldots, K\}$, where K is the number of states of variable i. (In general, K may depend on i, but we rarely write K_i). We write $X_{ni} = k$ if the i'th variable is in state k, where $k \in 1 : K$. Sometimes you will see a 1-of-K encoding, where $X_{ni} \in \{0, 1\}^K$, where $X_{nik} = I(X_{ni} = k)$. We also use j to index states, mostly of variables that are "parents" of X_i.

3 Bernoullis and multinomials

- We define the Bernoulli distribution $X \sim Be(\theta)$ for $X \in \{0, 1\}$ by

$$Be(X|\theta) = \theta^X (1 - \theta)^{1-X} \tag{1}$$

  We denote the minimal sufficient statistics for a Bernoulli distribution by the number of heads and tails: $N_1 = \sum_n I(X_n = 1)$, $N_0 = \sum_n I(X_n = 0)$. Alternatively, we can use N_1 and $N = N_1 + N_0$.

- We define the multinomial distribution $X \sim Mu(\theta)$ for $X \in \{1, \ldots, K\}$ by

$$Mu(X|\theta) = \prod_{j=1}^{K} \theta_j^{I(X=j)} \tag{2}$$

  Put another way, $p(X = j|\theta) = \theta_j$. We denote the sufficient statistics for a multinomial distribution by $N_j = \sum_n I(X_n = j)$.

- We define the Beta distribution $\theta \sim Beta(\alpha_1, \alpha_0)$ for $\theta \in [0, 1]$ by

$$\text{Beta}(\theta|\alpha_1, \alpha_0) = \frac{\Gamma(\alpha_1 + \alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_0)} \theta^{\alpha_1 - 1} (1 - \theta)^{\alpha_0 - 1} \tag{3}$$

  where $\Gamma(x)$ is the gamma function. Here $\alpha_0, \alpha_1 \in \mathbb{R}^+$ are called hyper parameters (pseudo counts) and $\alpha = \alpha_0 + \alpha_1$ is the equivalent sample size (strenght) of the prior.

- We define the Dirichlet distribution $\theta \sim Dir(\alpha_1, \ldots, \alpha_K)$ for $\theta \in [0, 1]^K$ by

$$\text{Dir}(\theta|\alpha_1, \ldots, \alpha_K) \propto \prod_{j=1}^{K} \theta_j^{\alpha_j - 1} \tag{4}$$

  Here $\alpha_j \in \mathbb{R}^+$ are called hyper parameters (pseudo counts) and $\alpha = \sum_j \alpha_j$ is the equivalent sample size.

- We define the likelihood as

$$L(\theta) = p(\mathcal{D}|\theta) \tag{5}$$

  and the log-likelihood as

$$\ell(\theta) = \log p(\mathcal{D}|\theta) \tag{6}$$

- We denote the maximum likelihood estimate by

$$\hat{\theta}^{ML} \in \arg\max_{\theta} p(\mathcal{D}|\theta) \tag{7}$$

  We denote the maximum a posterior estimate by

$$\hat{\theta}^{MAP} \in \arg\max_{\theta} p(\mathcal{D}|\theta) p(\theta) \tag{8}$$

  We denote the posterior mean estimate by

$$\hat{\theta}^{mean} = E[\theta|\mathcal{D}] \tag{9}$$

4 Naive Bayes classifier

- The 1d Gaussian density is denoted $\mathcal{N}(x|\mu, \sigma)$.
- In a generative classifer, the class prior is usually denoted $p(Y = c) = \pi_c$, if we assume Y has a multinomial distribution.
- In the naive Bayes model, we have

$$p(x|y = c) = \prod_{i=1}^{D} p(x_i|y = c) \tag{10}$$

  In the case of K-ary features, we have

$$p(x_i|y = c) = \prod_{k} \theta_{ick}^{I(X_i=k)} \tag{11}$$

  where $\theta_{ick} = P(X_i = k|Y = c)$. The sufficient statistics are N_{ick}, which is the number of times $X_i = k$ amongst those training cases where $Y = c$. In the case of binary features, we have

$$p(x_i|y = c) = \theta_{ic}^{I(X_i=1)}(1 - \theta_{ic})^{I(X_i=0)} \tag{12}$$

  where $\theta_{ic} = P(X_i = 1|Y = c)$. The sufficient statistics are N_{ic1}, the number of times $X_i = 1$ amongst cases where $Y = c$, and $N_{ic} = N_c$, the number of times $X_i = 0$ or $X_i = 1$ in cases where $Y = c$.

5 Markov chains

- The transition matrix is $T_{jk}^i = p(X_i = k|X_{i-1} = j)$, which is independent of i if the chain is stationary. The sufficient statistics to estimate this are the observed number of $j \to k$ transitions: $N_{jk} = \sum_{n=1}^{N} \sum_{i=2}^{D} I(X_{ni} = k, X_{ni-1} = j)$. There is no i index since we assume the parameters are shared (tied) across time.
- The initial state distribution is $\pi_k^1 = p(X_1 = k)$.
- The stationary distribution is π which satisfies $\pi T = \pi$ (if we treat π as a row vector).

6 Information theory

- The entropy of a random variable $X \in 1 : K$ with discrete distribution p is denoted by

$$H(p) = H(X) = -\sum_{k=1}^{K} p(X = k) \log_2 p(X = k) = -\sum_{k} p_k \log p_k \tag{13}$$

  The joint entropy is denoted $H(X, Y)$ and the conditional entropy as $H(X|Y)$. The mutual information is denoted $I(X, Y)$ (often written as $I(X; Y)$). The Kullback-Leibler divergence between two distributions is denoted $KL(p||q)$.