**E0 270 Machine Learning** **Lecture 3 (Jan 15, 2013)**

# Discriminative Probabilistic Models for Classification

*Lecturer: Shivani Agarwal*

**Disclaimer:** These notes are a *brief* summary of the topics covered in the lecture. They are not a substitute for the full lecture.

## Outline

- Logistic regression
- Other 'link' functions
- Multiclass extensions
- Loss minimization view

## 1 Logistic Regression

Consider first binary classification with label space $\mathcal{Y} = \{\pm 1\}$. We have seen that the Bayes optimal classifier depends only on the conditional class probabilities determined by $\eta(x)$. Therefore, if the goal is simply to learn an accurate classifier, one need not model the full joint distribution $D$, which is what the generative models effectively do; instead, it is sufficient to model only $\eta(x)$. One approach for this is to assume a parametric form for $\eta(x)$, and then estimate the parameters from the training sample using maximum likelihood estimation.

Let us write

$$\eta(x) = \mathbf{P}(y = 1|x) = \frac{1}{1 + e^{-f(x)}}$$

for some function $f : \mathcal{X} \to \mathbb{R}$.[1] Clearly, a Bayes optimal classifier can then be written as

$$h^*(x) = \text{sign}\Big(\frac{1}{1 + e^{-f(x)}} - \frac{1}{2}\Big) = \text{sign}(1 - e^{-f(x)}) = \text{sign}(f(x))\,.$$

So far, we have not made any assumptions on $\eta$, as there always exists a function $f : \mathcal{X} \to \mathbb{R}$ given by $f(x) = \ln\big(\frac{\eta(x)}{1-\eta(x)}\big)$ for which the above holds. However, this re-writing now allows us to consider parameterizing $f : \mathcal{X} \to \mathbb{R}$ instead of $\eta : \mathcal{X} \to [0, 1]$ directly. Note also that this gives

$$1 - \eta(x) = \mathbf{P}(y = -1|x) = 1 - \frac{1}{1 + e^{-f(x)}} = \frac{1}{1 + e^{f(x)}}\,,$$

so that we can write for each $y \in \{\pm 1\}$,

$$\mathbf{P}(y|x) = \frac{1}{1 + e^{-yf(x)}}\,.$$

[1] It can be verified that for the case of multivariate normal class-conditional densities on $\mathbb{R}^d$ with shared covariance matrix, as well as the case of conditionally independent features in $\{0, 1\}^d$, $\eta(\mathbf{x})$ has this form with $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ for some $\mathbf{w}, b$ that depend on the distribution parameters. In fact this holds for a larger class of settings in which the class-conditional distributions come from a certain form of **exponential family** model.

Now, let $\mathcal{X} = \mathbb{R}^d$, and assume $\eta(\mathbf{x})$ can be written in the above form for a linear function $f : \mathcal{X} \to \mathbb{R}$ of the form[2]

$$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$$

for some $\mathbf{w} \in \mathbb{R}^d$:

$$\eta(\mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^\top \mathbf{x}}} \,.$$

The parameter vector $\mathbf{w}$ can then be estimated from the training sample $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$, for example using maximum likelihood estimation. Specifically, the (conditional) likelihood of $\mathbf{w}$ is given by

$$\begin{aligned} \mathcal{L}(\mathbf{w}) &= \mathbf{P}\big(y_1, \ldots, y_m | \mathbf{x}_1, \ldots, \mathbf{x}_m; \mathbf{w}\big) \\ &= \prod_{i=1}^{m} \mathbf{P}\big(y_i | \mathbf{x}_i; \mathbf{w}\big) \\ &= \prod_{i=1}^{m} \frac{1}{1 + e^{-y_i \mathbf{w}^\top \mathbf{x}_i}} \,, \end{aligned}$$

and the log-likelihood is therefore given by

$$\ln \mathcal{L}(\mathbf{w}) = -\sum_{i=1}^{m} \ln\left(1 + e^{-y_i \mathbf{w}^\top \mathbf{x}_i}\right).$$

Unfortunately there is no closed-form expression for the parameters $\widehat{\mathbf{w}}$ maximizing the above log-likelihood, but these can be found using numerical optimization methods. The resulting plug-in classifier given by

$$h_S(\mathbf{x}) = \operatorname{sign}\big(\widehat{\mathbf{w}}^\top \mathbf{x}\big)$$

is termed the **(linear) logistic regression** classifier.[3]

**Exercise.** Let $\widehat{\eta}_S(x)$ denote an estimate of the conditional class probability function $\eta(x)$. Show that the **0-1 regret** of the corresponding plug-in classifier $h_S(x) = \operatorname{sign}(\widehat{\eta}_S(x) - \frac{1}{2})$, i.e. the difference of its 0-1 error from the Bayes error, is bounded by twice the expected absolute difference between $\widehat{\eta}_S(x)$ and $\eta(x)$:

$$\mathrm{er}_D^{0\text{-}1}[h_S] - \mathrm{er}_D^{0\text{-}1,*} \;\leq\; 2\,\mathbf{E}_{x\sim\mu}\big[\big|\widehat{\eta}_S(x) - \eta(x)\big|\big] \,.$$

## 2 Other 'Link' Functions

In the above derivation of the logistic regression classifier, we implicitly used the **logistic link function** $\psi : [0,1] \to \mathbb{R}$ defined as[4]

$$\psi(\eta) = \ln\left(\frac{\eta}{1-\eta}\right),$$

with corresponding **inverse logistic link function** $\psi^{-1} : \mathbb{R} \to [0,1]$ given by

$$\psi^{-1}(f) = \frac{1}{1 + e^{-f}} \,,$$

which is also called the **logistic sigmoid function** and was used to map real-valued numbers $f(x)$ to class probabilities $\eta(x)$. The main critical property of such a link function $\psi : [0,1] \to \mathbb{R}$ is that it needs to be surjective and strictly increasing (and therefore invertible on $\mathbb{R}$), and one can potentially use other link

[2] Note that one can accommodate a function of the form $\mathbf{w}^\top \mathbf{x} + b$ by augmenting the feature vector $\mathbf{x}$ with an additional constant component to create $\mathbf{x}' = (\mathbf{x}, 1) \in \mathbb{R}^{d+1}$, and taking $\mathbf{w}' = (\mathbf{w}, b) \in \mathbb{R}^{d+1}$, so that $\mathbf{w}'^\top \mathbf{x}' = \mathbf{w}^\top \mathbf{x} + b$.

[3] Often, in practice, one assumes a prior distribution over the parameters $\mathbf{w}$, such as a multivariate normal or Laplace distribution, and finds a **maximum a posteriori (MAP)** estimate $\widehat{\mathbf{w}}$. We will look at the effect of using such MAP estimation in later lectures. A fully Bayesian treatment where one uses the full posterior distribution over $\mathbf{w}$ for classification leads to **Bayesian logistic regression**. If in addition one assumes a prior distribution over the parameters of the prior on $\mathbf{w}$, this leads to a **hierarchical Bayesian** treatment of logistic regression.

[4] Note that we overload notation here by using $\eta$ and $f$ to denote numbers in $[0,1]$ and $\mathbb{R}$, respectively, rather than functions; the usage should be clear from context.

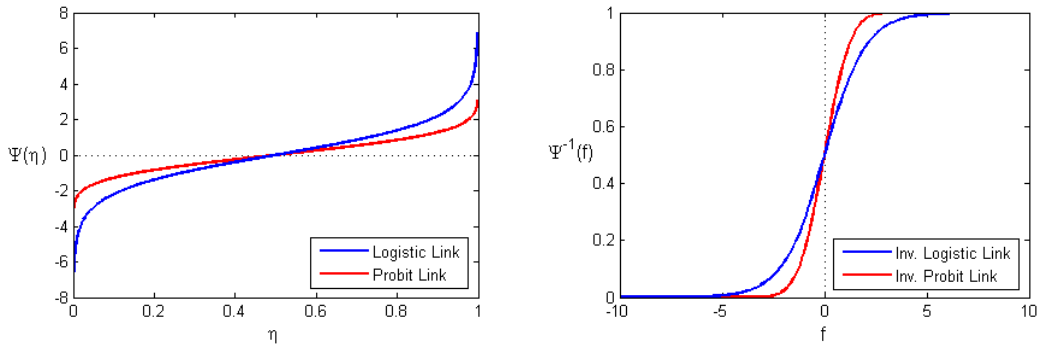


Figure 1: Logistic and probit link functions and their inverses.

functions satisfying these properties as well. In particular, the cumulative distribution function (CDF) of any continuous random variable that has a positive density over all of $\mathbb{R}$ (and is therefore a strictly increasing surjective function from $\mathbb{R}$ to $[0,1]$) is a candidate for the inverse link function; equivalently, the inverse of any such CDF is a candidate for the link function. For example, the **probit link function** uses the inverse of the standard normal CDF $\Phi$:

$$\psi(\eta) = \Phi^{-1}(\eta)\,, \quad \text{where } \Phi(f) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{f} e^{-t^2/2}\, dt\,;$$

using this link function gives rise to what is called **probit regression**.

Both the logistic and probit links and their inverses are shown in Figure 1. Other link functions have also been used as a basis for designing similar discriminative probabilistic models for binary classification. Can you think of some reasons for why one might want to choose one link function over another?

## 3 Multiclass Extensions

Consider now multiclass classification with $r > 2$ classes: $\mathcal{Y} = \{1, \ldots, r\}$. A discriminative probabilistic model in this case will attempt to model the conditional class probabilities $\eta_y(x) = \mathbf{P}(y|x)$ for all $y \in [r]$.

In the binary case, we parametrized the class probability function $\eta : \mathcal{X} \to [0,1]$ by parametrizing a real-valued function $f : \mathcal{X} \to \mathbb{R}$ and using an inverse link function $\psi^{-1} : \mathbb{R} \to [0,1]$ to map back to a probability estimate: $\eta(x) = \psi^{-1}(f(x))$. In the multiclass case, we have a vector-valued class probability function $\boldsymbol{\eta} : \mathcal{X} \to \Delta_r$, where $\Delta_r = \left\{\boldsymbol{\eta} \in \mathbb{R}^r : \eta_y \geq 0 \;\forall y \in [r], \sum_{y=1}^{r} \eta_y = 1\right\}$ denotes the probability simplex of $r$-dimensional probability vectors. In this case, we can parametrize $\boldsymbol{\eta} : \mathcal{X} \to \Delta_r$ by parametrizing a vector-valued function $\mathbf{f} : \mathcal{X} \to \mathbb{R}^{r-1}$ and using a (invertible) **multiclass or multinomial link function** $\boldsymbol{\psi} : \Delta_r \to \mathbb{R}^{r-1}$, or rather its inverse, $\boldsymbol{\psi}^{-1} : \mathbb{R}^{r-1} \to \Delta_r$, to map back to a probability vector: $\boldsymbol{\eta}(x) = \boldsymbol{\psi}^{-1}(\mathbf{f}(x))$.

A widely used multinomial link function is the **multinomial logit link**, whose inverse $\boldsymbol{\psi}^{-1} : \mathbb{R}^{r-1} \to \Delta_r$ is given by

$$(\boldsymbol{\psi}^{-1}(\mathbf{f}))_y = \begin{cases} \frac{\exp(f_y)}{1+\sum_{y'=1}^{r-1} \exp(f_{y'})} & \text{if } y \in [r-1] \\ \frac{1}{1+\sum_{y'=1}^{r-1} \exp(f_{y'})} & \text{if } y = r. \end{cases}$$

For example, let $\mathcal{X} = \mathbb{R}^d$, and assume $\boldsymbol{\eta}(\mathbf{x})$ can be written as

$$\eta_y(\mathbf{x}) = \begin{cases} \frac{\exp(f_y(\mathbf{x}))}{1+\sum_{y'=1}^{r-1} \exp(f_{y'}(\mathbf{x}))} & \text{if } y \in [r-1] \\ \frac{1}{1+\sum_{y'=1}^{r-1} \exp(f_{y'}(\mathbf{x}))} & \text{if } y = r \end{cases}$$

for some linear function $\mathbf{f} : \mathcal{X} \to \mathbb{R}^{r-1}$ of the form

$$\mathbf{f}(\mathbf{x}) = \mathbf{W}^\top \mathbf{x}$$

for some $\mathbf{W} \in \mathbb{R}^{d\times(r-1)}$. As before, one can find a maximum likelihood (or other) estimate of the parameter matrix $\mathbf{W}$ from the training data $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$, use the estimated parameters $\widehat{\mathbf{W}}$ to construct a plug-in class probability estimate $\widehat{\boldsymbol{\eta}}_S(\mathbf{x})$, and then classify according to

$$h_S(\mathbf{x}) \;\in\; \arg\max_{y\in[r]} \left(\widehat{\boldsymbol{\eta}}_S(\mathbf{x})\right)_y .$$

In practice, one often uses a 'redundant' parametrization of $\boldsymbol{\eta}(\mathbf{x})$ in terms of an $r$-dimensional vector function $\mathbf{f} : \mathcal{X}\to\mathbb{R}^r$, for example using the **softmax function** which maps vectors in $\mathbb{R}^r$ to $\Delta_r$:[5]

$$\eta_y(\mathbf{x}) = \frac{\exp(f_y(\mathbf{x}))}{\sum_{y'=1}^{r}\exp(f_{y'}(\mathbf{x}))} .$$

Taking again the vector function $\mathbf{f} : \mathcal{X}\to\mathbb{R}^r$ to be a linear function of the form

$$\mathbf{f}(\mathbf{x}) = \mathbf{W}^\top\mathbf{x}$$

for some $\mathbf{W} \in \mathbb{R}^{d\times r}$, one can again use a maximum likelihood (or other) estimate $\widehat{\mathbf{W}}$ to construct a plug-in estimate $\widehat{\boldsymbol{\eta}}_S(\mathbf{x})$; in this case, the plug-in classifier can be seen to be of the form

$$h_S(\mathbf{x}) \;\in\; \arg\max_{y\in[r]} \widehat{\mathbf{w}}_y^\top\mathbf{x} ,$$

where $\widehat{\mathbf{w}}_y$ denotes the $y$-th column of $\widehat{\mathbf{W}}$. The softmax parametrization can be seen to be equivalent to the multinomial logit link parametrization up to scaling factors. Both approaches generalize the logistic link function parametrization for binary classification, and the resulting classifiers are referred to as **(linear) multiclass/multinomial logistic regression** classifiers.

# 4 Loss Minimization View

Let us briefly look at an alternative view of the linear logistic regression classifier described earlier for binary classification, with $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{\pm 1\}$. Clearly, among all linear classifiers of the form $h_{\mathbf{w}}(\mathbf{x}) = \text{sign}(\mathbf{w}^\top\mathbf{x})$, the ideal classifier would be one that minimizes the 0-1 error w.r.t. $D$:

$$\mathbf{w}^* \;\in\; \arg\min_{\mathbf{w}\in\mathbb{R}^d} \text{er}_D^{0\text{-}1}[h_{\mathbf{w}}] .$$

Since $D$ is unknown, one might look for a parameter vector that minimizes the **empirical 0-1 error** on the training sample $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m))$ instead[6], defined for any classifier $h : \mathcal{X}\to\{\pm 1\}$ as

$$\widehat{\text{er}}_S^{0\text{-}1}[h] = \frac{1}{m}\sum_{i=1}^{m}\mathbf{1}\big(h(\mathbf{x}_i) \neq y_i\big) .$$

Minimizing this 0-1 empirical error (over linear classifiers) would give

$$\mathbf{w}_S^* \;\in\; \arg\min_{\mathbf{w}\in\mathbb{R}^d} \widehat{\text{er}}_S^{0\text{-}1}[h_{\mathbf{w}}] = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{m}\sum_{i=1}^{m}\mathbf{1}\big(\text{sign}(\mathbf{w}^\top\mathbf{x}_i) \neq y_i\big) .$$

Unfortunately, this turns out to be a computationally difficult optimization problem due to the discrete indicator function (in fact it is NP-hard to solve). Consequently, one often minimizes the empirical error on $S$ w.r.t. some other 'surrogate' loss function $\ell : \{\pm 1\}\times\mathbb{R}\to\mathbb{R}_+$ that is a continuous and sufficiently smooth approximation to the 0-1 loss to allow for efficient minimization; often, it is desirable to have the loss be convex in its second argument. In particular, define the **logistic loss** $\ell_{\log} : \{\pm 1\}\times\mathbb{R}\to\mathbb{R}_+$ as

$$\ell_{\log}(y, f) = \log_2(1 + e^{-yf}) .$$

[5] In the multiclass case, multivariate normal class-conditional densities in $\mathbb{R}^d$ with shared covariance matrix as well as conditionally independent features in $\{0,1\}^d$ both lead to $\eta_y(\mathbf{x})$ of this form with $f_y(\mathbf{x}) = \mathbf{w}_y^\top\mathbf{x} + b_y$ for some $\mathbf{w}_y, b_y$ (again, this holds for a wider class of class-conditional distributions that come from a certain form of exponential family model).

[6] Since $S$ contains examples drawn i.i.d. from $D$, minimizing the error on $S$ seems intuitively to be a reasonable thing to do; we will see formal reasons later for why this makes sense.

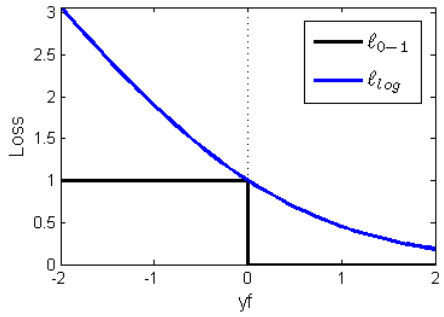


Figure 2: Logistic and 0-1 losses, as a function of the margin $yf$.

Note also that for real-valued predictions $f \in \mathbb{R}$ used for classification via $\text{sign}(f)$, the 0-1 loss becomes $\ell_{0\text{-}1} : \{\pm 1\} \times \mathbb{R} \to \{0, 1\}$, defined as

$$\ell_{0\text{-}1}(y, f) = \mathbf{1}(\text{sign}(f) \neq y) = \mathbf{1}(yf < 0)\,.$$

Figure 2 shows plots of both the logistic loss and the 0-1 loss as a function of the **margin** $yf$. As can be seen, the logistic loss is convex in its second argument and forms an upper bound on the 0-1 loss. We can define the **empirical logistic error** of a real-valued function $f : \mathcal{X} \to \mathbb{R}$ on the training sample $S$ as

$$\widehat{\text{er}}_S^{\log}[f] = \frac{1}{m} \sum_{i=1}^{m} \log_2(1 + e^{-y_i f(\mathbf{x}_i)})\,.$$

Minimizing this empirical logistic error over all linear functions of the form $f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ then gives

$$\mathbf{w}_S \in \arg\min_{\mathbf{w} \in \mathbb{R}^d} \widehat{\text{er}}_S^{\log}[f_{\mathbf{w}}] = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{m} \sum_{i=1}^{m} \log_2(1 + e^{-y_i \mathbf{w}^\top \mathbf{x}_i})\,.$$

This yields the same solution as the linear logistic regression classifier! This gives an alternative view of logistic regression that makes no assumptions directly on the conditional probability function $\eta$, but rather simply minimizes the empirical logistic error on the training sample over some class of functions (in this case linear functions). This view has been helpful in understanding statistical consistency properties of logistic regression classifiers in recent years. We will see more examples of such loss minimization algorithms, often called **empirical ($\ell$-)risk minimization (ERM or $\ell$-ERM)** algorithms (where $\ell$ denotes the loss being minimized), in the coming lectures.

**Acknowledgments.** Thanks to Harikrishna Narasimhan for help in preparing the plots for Figures 1–2.