IFT 6269: Probabilistic Graphical Models **Fall 2016**

Lecture 6 — September 20

Lecturer: Simon Lacoste-Julien *Scribe: Zakaria Soliman*

Disclaimer: These notes have only been lightly proofread.

6.1 Linear Regression

6.1.1 Motivation

We want to learn a prediction function $f : \mathcal{X} \to \mathcal{Y}$. Where $\mathcal{X} \subseteq \mathbb{R}^d$ and if:

(1) $\mathcal{Y} = \{0, 1\}$, it's a binary classification

(2) $\mathcal{Y} = \{0, 1, \cdots, k\}$, it's a multiclass classification

(3) $\mathcal{Y} \subseteq \mathbb{R}$, it's a regression problem.

There are several perspectives in modeling the distribution of the data:

generative perspective

Here, we model the joint distribution $p(x, y)$. We make more assumptions in this case. This leads it to be less robust for predictions (but is a more flexible approach if we are not sure what is the task we are trying to solve).

conditional perspective

We only model the conditional probability $p(y|x)$. Early 2000s, it was called the discriminative perspective, but Simon prefers to refer to it now as the conditional approach.

fully discriminative perspective

Models $f : \mathcal{X} \to \mathcal{Y}$ directly and estimate the function $\hat{f}$ by using the loss $\ell(y, y')$ information. This approach is the most robust.

6.1.2 Linear regression model

We take a conditional approach to regression. Let $Y \in \mathbb{R}$ and let's assume that Y depends linearly on $X \in R^d$. Linear regression is a model of the following form:

$$p(y|\mathbf{x}, \mathbf{w}) = \mathcal{N}(y|\langle \mathbf{w}, \mathbf{x} \rangle, \sigma^2)$$

Where $\mathbf{w} \in \mathbb{R}^d$ is the parameter (or weight) vector. Equivalently, we could also rewrite the model as

$$Y = \mathbf{w}^\top \mathbf{X} + \epsilon$$

Where the noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is a random variable that is independent of X

Remark 6.1.1 *Note that if there is an offset $w_0 \in \mathbb{R}$, that is, if $Y = w_0 + \mathbf{w}^\top X + \epsilon$, we will use an "offset" notation for $\mathbf{x}$:*

$$\mathbf{x} = \begin{pmatrix} \tilde{\mathbf{x}} \\ 1 \end{pmatrix},$$

where $\tilde{\mathbf{x}} \in \mathbb{R}^{d-1}$ and 1 is the constant feature. Thus, we have:

$$\mathbf{w}^\top \mathbf{x} = \mathbf{w}_{1:d-1}^\top \tilde{\mathbf{x}} + w_d$$

Where w_d is the bias/offset

Let $D = (\mathbf{x}_n, y_n)_{i=1}^n$ be a training set of conditionally i.i.d. random variables i.e. $X_i \sim$ whatever and $Y_i|X_i \sim \mathcal{N}(\langle \mathbf{w}\,, X_i \rangle, \sigma^2)$. Each y_i is a response on observation $\mathbf{x}_i$. We consider the conditional likelihood of all outputs given all inputs:

$$p(y_1, \cdots, y_n | \mathbf{x}_1, \cdots, \mathbf{x}_n; \mathbf{w}, \sigma^2) = \prod_{i=1}^n p(y_i | \mathbf{x}_i; \mathbf{w}, \sigma^2).$$

And we have that $Y_i|X_i \overset{\text{indep}}{\sim} \mathcal{N}(\mathbf{w}^\top X_i\,, \sigma^2)$ (i.e. $p(y_i|\mathbf{x}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{2\sigma^2}\right)$) taking the log-likelihood gives us the following expression:

$$\begin{aligned} \log p(y_{1:n} | \mathbf{x}_{1:n}; \mathbf{w}, \sigma^2) &= \sum_{i=1}^n \log p(y_i|\mathbf{x}_i) \\ &= \sum_{i=1}^n \left[-\frac{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) \right] \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^n \frac{(y_i - \mathbf{w}^\top \mathbf{x}_i)^2}{\sigma^2}. \end{aligned}$$

Notice that maximizing the likelihood comes down to the following minimization problem w.r.t. $\mathbf{w}$:

$$\text{find} \quad \hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \textstyle\sum_{i=1}^n (y_i - \mathbf{w}^\top \mathbf{x}_i)^2.$$

Define the design matrix $\mathbf{X}$ as

$$\mathbf{X} = \begin{pmatrix} \text{——} \; \mathbf{x}_1^\top \; \text{——} \\ \vdots \\ \text{——} \; \mathbf{x}_n^\top \; \text{——} \end{pmatrix} \in \mathbb{R}^{n \times d}$$

and denote by $\mathbf{y}$ the vector of coordinates $\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$. This notation allows us to rewrite the residual sum of squares in a more compact fashion as:

$$\sum_{i=1}^{n}(y_i - \mathbf{w}^\top \mathbf{x}_i)^2 = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2$$

Thus, we can rewrite the log likelihood as:

$$-\log p(\mathbf{y}|\mathbf{x}) = \frac{\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2}{2\sigma^2} + \frac{n}{2}\log(2\pi\sigma^2)$$

Finally, the minimization problem over $\mathbf{w}$ can be rewritten as:

$$\text{find} \quad \hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2.$$

Remark 6.1.2 *The minimization of $\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2$ w.r.t. $\mathbf{w}$ can also be viewed geometrically as choosing $\hat{\mathbf{w}}$ so that the vector $\mathbf{X}\hat{\mathbf{w}}$ is the orthogonal projection of $\mathbf{y}$ onto the column space of $\mathbf{X}$*

Now let us find $\hat{\mathbf{w}}$:

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{w}}(\mathbf{y} - \mathbf{X}\mathbf{w})^\top(\mathbf{y} - \mathbf{X}\mathbf{w}) &= \frac{\partial}{\partial \mathbf{w}}\left[\|\mathbf{y}\|^2 - 2\mathbf{y}^\top\mathbf{X}\mathbf{w} + \mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w}\right] \\
&= 0 - 2\mathbf{X}^\top\mathbf{y} + 2\mathbf{X}^\top\mathbf{X}\mathbf{w} = 0 && (\text{using } \nabla_{\mathbf{w}}(\mathbf{w}^\top\mathbf{A}\mathbf{w}) = (\mathbf{A} + \mathbf{A}^\top)\mathbf{w})) \\
&\iff \boxed{(\mathbf{X}^\top\mathbf{X})\mathbf{w} = \mathbf{X}^\top\mathbf{y}} && \text{normal equation}
\end{aligned}$$

- If $\mathbf{X}^\top\mathbf{X}$ is invertible, there is a unique solution $\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$
- If $n < d$, then $\mathbf{X}$ is not full rank and so $\mathbf{X}^\top\mathbf{X}$ is not invertible. In this case we could use the pseudo-inverse of $\mathbf{X}$, $\mathbf{X}^\dagger$ and choose the minimum norm $\|\mathbf{w}\|$ solution amongst $\arg\min_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2$. The problem we face is that the pseudo-inverse is not numerically stable.

In the latter case, it would be better to use regularization techniques (see next section).

6.1.3 Ridge regression

We can either interpret ridge regression as adding a norm regularizer to the least-square EMR, or as replacing the MLE for $\mathbf{w}$ with a MAP by adding a prior $p(\mathbf{w})$:

$$\log p(\mathbf{w}|\mathbf{y},\mathbf{x}) = \log p(y_{1:n}|\mathbf{x}_{1:n};\mathbf{w}) + \log p(\mathbf{w}) + cst$$

Where $p(\mathbf{w})$ is the prior over $\mathbf{w}$ and:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|0, \frac{\mathbf{I}}{\lambda})$$

So we have that:

$$\log p(\mathbf{w}|\mathbf{y},\mathbf{x}) = \log p(y_{1:n}|\mathbf{x}_{1:n};\mathbf{w}) + cst - \frac{\lambda}{2}\left\|\mathbf{w}\right\|^2$$

and then,

$$\begin{aligned}\nabla_{\mathbf{w}} = 0 &\Rightarrow (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})\mathbf{w} = \mathbf{X}^\top\mathbf{y} \\ &\Rightarrow \hat{\mathbf{w}}_{MAP} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}\end{aligned}$$

Notice that $(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})$ is always invertible.

Remark 6.1.3 $-\log p(\mathbf{w}|\mathbf{y},\mathbf{x})$ *is strongly convex in* $\mathbf{w}$*. So there is a unique global minimum*

Remark 6.1.4 *It is good practice to standardize or normalize the features. Standardizing means make the features have empirical zero mean and unit standard deviation; normalizing can mean different things, e.g. scale them to* $[0,1]$ *or to a unit norm.*

6.2 Logistic Regression

Let's turn our attention to classification problems. For this model, we will assume that $Y \in \{0,1\}$ and $X \in \mathbb{R}^d$. We make no additional assumptions apart that $p(\mathbf{x}|Y=1)$ and $p(\mathbf{x}|Y=0)$ are densities. Our goal is to model $p(Y|X)$

$$\begin{aligned}p(Y=1|X=\mathbf{x}) &= \frac{p(Y=1,X=\mathbf{x})}{p(Y=1,X=\mathbf{x}) + p(Y=0,X=\mathbf{x})} \\ &= \frac{1}{1+\frac{p(Y=1,X=\mathbf{x})}{p(Y=0,X=\mathbf{x})}} \\ &= \frac{1}{1+\exp(-f(\mathbf{x}))}\end{aligned}$$

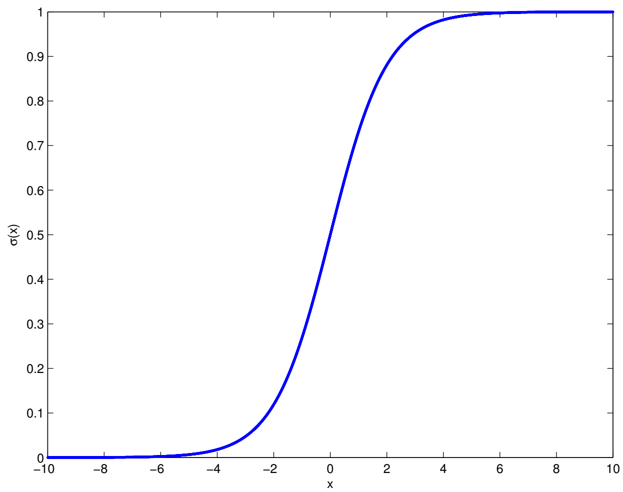

Figure 6.1: Sigmoid function.

Where

$$f(x) = \log \underbrace{\frac{p(X = \mathbf{x}|Y = 1)}{p(X = \mathbf{x}|Y = 0)}}_{\text{class-conditional ratio}} + \log \underbrace{\frac{p(Y = 1)}{p(Y = 0)}}_{\text{prior odd ratio}}$$

Is the log odds ratio. In general we have:

$$p(Y = 1|X = \mathbf{x}) = \sigma(f(\mathbf{x}))$$

where $\sigma(z) := \frac{1}{1+e^{-z}}$ is the sigmoid function shown in Figure 2.1.

The sigmoid function has the following properties:

Property 6.2.1

$$\forall z \in \mathbb{R}, \sigma(-z) = 1 - \sigma(z)$$

Property 6.2.2

$$\forall z \in \mathbb{R}, \sigma'(z) = \sigma(z)(1 - \sigma(z)) = \sigma(z)\sigma(-z)$$

Example 6.2.1 *Finally, we make the following observation that a very large class of probabilistic models yield logistic-regression types of models (thus explaining why logistic regression is fairly robust).*

Consider that the class conditional is in the exponential family:

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x}) \exp(\boldsymbol{\eta}^\top \mathbf{T}(\mathbf{x}) - A(\boldsymbol{\eta})).$$

$$\begin{aligned} f(\mathbf{x}) &= \log \frac{p(X = \mathbf{x}|Y = 1)}{p(X = \mathbf{x}|Y = 0)} + \log \frac{p(Y = 1)}{p(Y = 0)} \\ &= (\boldsymbol{\eta}_1 - \boldsymbol{\eta}_0)^\top \mathbf{T}(\mathbf{x}) + A(\boldsymbol{\eta}_0) - A(\boldsymbol{\eta}_1) + \log(\frac{\pi}{1-\pi}) \\ &= \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}) \end{aligned}$$

Where $\mathbf{w} = \binom{\boldsymbol{\eta}_1 - \boldsymbol{\eta}_0}{A(\boldsymbol{\eta}_0) - A(\boldsymbol{\eta}_1) + \log(\frac{\pi}{1-\pi})}$ *and* $\boldsymbol{\phi}(\mathbf{x}) = \binom{\mathbf{T}(\mathbf{x})}{1}$*. Thus we have a logistic regression model with features* $\boldsymbol{\phi}(\mathbf{x})$*:*

$$p(y = 1|\mathbf{x}) = \sigma(\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}))$$